Discovering Perceptions in Online Social Media: A Probabilistic Approach

Derek Doran  
*Wright State University - Main Campus*, derek.doran@wright.edu

Swapna S. Gokhale

Aldo Dagnino

Follow this and additional works at: [https://corescholar.libraries.wright.edu/knoesis](https://corescholar.libraries.wright.edu/knoesis)

Part of the Bioinformatics Commons, Communication Technology and New Media Commons, Databases and Information Systems Commons, OS and Networks Commons, and the Science and Technology Studies Commons

Repository Citation  
[https://corescholar.libraries.wright.edu/knoesis/1047](https://corescholar.libraries.wright.edu/knoesis/1047)

This Article is brought to you for free and open access by the The Ohio Center of Excellence in Knowledge-Enabled Computing (Kno.e.sis) at CORE Scholar. It has been accepted for inclusion in Kno.e.sis Publications by an authorized administrator of CORE Scholar. For more information, please contact library-corescholar@wright.edu.
DISCOVERING PERCEPTIONS IN ONLINE SOCIAL MEDIA: A PROBABILISTIC APPROACH*

DEREK DORAN  
Department of Computer Science & Engineering  
Kno.e.sis Research Center  
Wright State University  
3640 Colonel Glenn Highway  
Dayton, OH, 45435, USA  
derek.doran@wright.edu

SWAPNA S. GOKHALE  
Department of Computer Science & Engineering  
University of Connecticut  
371 Fairfield Way Unit 4155  
Storrs, CT, 06269, USA  
ssg@engr.uconn.edu

ALDO DAGNINO  
Industrial Software Systems, Data Analytics  
ABB Corporate Research  
940 Main Campus Drive  
Raleigh, NC, 27603, USA  
aldo.dagnino@us.abb.com

People across the world habitually turn to online social media to share their experiences, thoughts, ideas, and opinions as they go about their daily lives. These posts collectively contain a wealth of insights into how masses perceive their surroundings. Therefore, extracting people’s perceptions from social media posts can provide valuable information about pertinent issues such as public transportation, emergency conditions, and even reactions to political actions or other activities. This paper proposes a novel approach to extract such perceptions from a corpus of social media posts originating from a given broad geographical region. The approach divides the broad region into a number of sub-regions, and trains language models over social media conversations within these sub-regions. Using Bayesian and geo-smoothing methods, the ensemble of language models can be queried with phrases embodying a perception. Discrete and continuous visualization methods represent the extent to which social media posts within the sub-regions express the query. The capabilities of the perception mining approach are illustrated using transportation-themed scenarios.

Keywords: Social media; Perception mining; Language models

*This is a pre-print of a paper that has been accepted for publication in the International Journal of Software Engineering & Knowledge Engineering.
1. Introduction and Motivation

Online social media services are now deeply rooted in our modern culture. Furthermore, widespread proliferation of mobile devices that are always-on and perpetually-connected provides easy access to these services. Such ubiquitous access enables people to publicly and instantly share their thoughts, opinions, feelings, and observations about the world that surrounds them. These expressions of people’s perceptions about their locale and environment can offer invaluable insights into the issues and concerns that they care about deeply. For example, the perceptions can provide a unique window into political crises such as the Arab Spring uprising in Egypt [36], the 2009 Iran Elections [7], and the 2011 Tunisian Uprising [53]. Finding these perceptions can also guide aid and recovery efforts following natural disasters; in fact, social media has earned widespread praise in the relief and recovery operations following the 2011 Tohoku earthquake and tsunami [17], the 2010 Haiti earthquake [52], and the 2013 Oklahoma tornadoes [10]. Perceptions about these extraordinary events, which impact broad swaths of people, can be easily extracted from social media posts containing rich contextual information by aggregating them according to the common hashtags and other topical keywords [44].

Our preliminary work [19] suggests that social media posts may also contain valuable perceptions that can rationalize or explain more casual and routine events, despite the fact that their impact is limited to a much smaller segment of the population. For example, perceptions expressing frustration about road construction, dysfunctional lights, and motorists’ accidents may offer important insights into the consistent traffic jams and delays that plague a local area. Perceptions may also provide a collective sense of the issues that affect day-to-day lives including local laws, shops, and public transit [18]. It is difficult, however, to extract such routine and local perceptions from social media posts because they discuss a wide variety of topics. Moreover, since many of these posts hold very few clues about context and location, it is even trickier to understand how these perceptions are influenced by the location or the surroundings from where they are shared. Such linking of “what” to “where” is instrumental in effectively utilizing these perceptions to remedy the concerns and frustrations they express. Fortunately, smartphones and mobile devices used to share these posts provide users with capabilities to geo-tag these posts or embed location information within them as metadata. As a result, a sizable number of posts [26] are now geo-tagged, which offers a unique opportunity to understand how the content or what in these posts is influenced by the location or from where they are shared.

Presently, perceptions about local or regional events can be extracted only using third party services such as Yelp\(^a\) or Foursquare\(^b\). However, these services can isolate locations only if the posts explicitly refer to specific businesses and landmarks.

\(^a\)http://www.yelp.com
\(^b\)http://www.foursquare.com
Opinions tied to individual businesses and landmarks, however, do not offer insights into an underlying abstract notion or a topic. For example, reviews about a specific fast food restaurant cannot indicate whether people in the area like to eat fast food, or eating fast food is popular. Instead, posts that talk about fast food generally, with or without reference to specific restaurants, provide clues about area-wide perceptions on the topic of fast food.

Semantic analysis is generally used to extract perceptions from a corpus of documents [21]. In theory, it may also be possible to analyze social media posts using semantic methods to extract people’s perceptions about a point of interest or an area [21]. However, state-of-the-art semantic models can only be trained using a small set of long, well-structured documents. Therefore, it is difficult to train these semantic models over social media posts because of their limited size (e.g. 140 characters on Twitter), and lack of rigor in following structural, linguistic rules [16]. Therefore, we hypothesize that a probabilistic model may be well-suited for mining perceptions from social media posts because such a model does not rely on structural regularities, and can instead find patterns in word usage and co-occurrences across a huge number of small sized, informal pieces of text.

In this paper, we propose a probabilistic approach to mine perceptions from social media posts, and associate them with their localities or neighborhoods. Partitioning a geographic region into non-overlapping sub-regions, the approach trains probabilistic language models over posts from these sub-regions. This ensemble of models is then queried with a phrase defining a topic-based perception to identify sub-regions where that perception runs strong. We illustrate our approach using Twitter feeds across four different areas of vastly different sizes, population densities, and other characteristics. Our illustrations, themed around transportation, focus on understanding perceptions about traffic, road construction, and mass transit. We verify our findings using publicly available statistics and service reports. Our illustrations thus show that despite regional diversities the approach can still pinpoint locations where a given perception runs strong.

The paper is organized as follows: Section 2 presents our methodology. Section 3 describes Twitter data. Section 4 applies the methodology to a series of illustrations themed around transportation. Section 5 compares related work. Conclusions and future directions are in Section 6.

2. Mining Perceptions

We define a “perception” as an observation noted by individuals that may be modulated by their personal feelings, beliefs, and perspectives. These personal feelings may be further influenced by the inherent regional characteristics. People’s perceptions about events and issues they encounter in their daily lives are often embodied in the words, terms and phrases that compose their spoken language. Virtually, there can exist an unlimited number of such topics and their associated perceptions; thus it is impossible to exhaustively define and model these possibilities. We
therefore propose a flexible approach where we divide a region into non-overlapping sub-regions and then for each individual sub-region we build a model to represent the general language used within its social media posts. We can query this ensemble of language models with a multi-word phrase that defines a perception for a specific topic. This querying will return the relative strength that the language of social media posts within each sub-region captures that specific perception. In the sequel, we describe the base language model, a geo-smoothing approach to enhance its performance, and graphical representations for visualizing the perceptions.

2.1. Model Representation

We divide a region $R$ into a $g \times g$ lattice, resulting in $g^2$ non-overlapping sub-regions of equal dimension. The language model of each sub-region $\ell_i$ represents the linguistic features of the social media posts that originate from the sub-region. It defines a probability distribution over $n$-grams, where an $n$-gram is an ordered sequence of $n$ words $(w_1, ..., w_n)$. Then, the maximum likelihood estimate of an $n$-gram, computed over a corpus of social media posts within $\ell_i \in \mathcal{L}$, is given by [3]:

$$P_{\ell_i}(w_1, ..., w_n) = \frac{c_{\ell_i}(w_1, ..., w_n)}{c_{\ell_i}(w_1, ..., w_{n-1})}$$  \hspace{1cm} (1)

where $c(.)$ is the number of times the sequence appears in the posts. The probability that a sub-region generates a phrase $T = (w_1, ..., w_k)$ is computed as the product of the probabilities of the $n$-grams that comprise $T$:

$$P(T|\ell_i) = \prod_{j=1}^{k-n+1} P_{\ell_i}(w_j, w_{j+1}, ..., w_{j+n-1})$$  \hspace{1cm} (2)

Larger values of $n$ allow the model to incorporate higher contextual information by considering higher order sequences of words. However, because social media posts are usually very short and carry limited information, the frequency with which progressively larger sequences of words appear across a corpus of such posts gets increasingly smaller. To combat this problem, prevalent language models for social media posts generally restrict themselves to unigrams, which only model distinct words independent of their order [11, 34]. While the distributions of unigrams can be estimated easily, we argue that they cannot adequately capture perceptions, which must be understood in the context of some topic or a subject. For example, unigrams trained over posts “I love driving” and “I hate driving” can only identify the concepts of “love” and “hate”, and cannot associate them with the topic of “driving”. By considering bigrams, however, we can identify how a subject loves or hates some object. Bigrams would thus allow us to identify that a given post features positive or negative sentiment related to driving, which is much richer information compared to unigrams.

The richness of information captured by bigrams makes them attractive as a basis of a language model. However, estimating their distribution is difficult due to
their short length and the unique, informal language used in social media posts. For example, it is likely that a sizable yet finite number of bigrams end with the word “York” (e.g. “New York”), so that it may be easy to estimate their relative frequencies. However, estimating the distribution of bigrams that end with an emoticon such as “:)” is difficult because an infinite number of such bigrams are not only possible, but may also appear in social media messages. We consider this tradeoff by representing the language of a sub-region as an interpolation of both bigram and unigram distribution of words. Through this interpolation, we compensate for the observed low count of a given bigram by incorporating the expected higher count of the unigram that completes it. For example, if the unigram “driving” is used frequently in a training corpus, we should expect that bigrams completed by this word (e.g. “hate driving”) are more likely to be seen even if such bigrams do not appear often. We thus define the probability of observing a bigram \((w_{j-1}, w_j)\) within a sub-region \(\ell_i\) as:

\[
P_{\ell_i}(w_{j-1}, w_j) = \lambda_1 \frac{c(w_{j-1}, w_j)}{c(w_{j-1})} + \lambda_2 \frac{c(w_j)}{|W(\ell_i)|}
\]

where \(\lambda_1 + \lambda_2 = 1\), \(|W(\ell_i)|\) is the number of distinct words in all the posts in \(\ell_i\) and \(c(w_j)/|W(\ell_i)|\) is the estimate of the unigram that completes the bigram [3].

We also incorporate smoothing in our model, which diverts some of the probability of bigrams seen frequently across a training corpus to those that have never been observed [12]. We employ the Modified Kneser-Ney (MKN) smoothing algorithm [31] because it has been shown to offer superior performance across interpolated language models such as our combined bigram/unigram model [12]. This algorithm diverts probability mass by subtracting a constant \(\hat{d}\) number of observations from the bigrams within a training corpus. To estimate the likelihood of a bigram \((w_{j-1}, w_j)\), the algorithm first estimates the probability of the unigram \(w_j\) that completes it by considering only the number of distinct bigrams \(w_j\) completes:

\[
P_c(w_j) = \frac{|\{w : c(w, w_j) > 0\}|}{\sum_v |\{w : c(w, v) > 0\}|}
\]

In other words, \(P_c(w_j)\) is the number of bigrams completed by \(w_j\) divided by the total number of unique bigrams in a corpus. This value is then weighted by the probability mass \(\lambda(w_{j-1})\) that was diverted by subtracting \(\hat{d}\) from the counts of known bigrams:

\[
\lambda(w_{j-1}) = \frac{\hat{d}|\{w : c(w_{j-1}, w) > 0\}|}{c(w_{j-1})}
\]

The probability of observing a bigram thus becomes:

\[
P_{\ell_i}(w_{j-1}, w_j) = \max\left(c(w_{j-1}, w_j) - \hat{d}, 0\right) \frac{c(w_{j-1})}{c(w_{j-1})} + \lambda(w_{j-1}) P_c(w_j)
\]
Note how if \((w_{j-1}, w_j)\) was never seen in the training corpus, its probability is given by \(\lambda \cdot \frac{c(w_j)}{|W(\ell)|}\) because under \(P_c(w_j)\), words that appear frequently but within few distinct contexts will not strongly influence the probability of the bigram. We select \(\hat{d}\) as the value that maximizes the log-likelihood that the model generates a given bigram:

\[
\hat{d} = \arg \max_d \sum_v c(v, w_j) \log P(\ell, v, w_j)
\]

This has a closed form approximation depending on whether \(c(w_{i-1}, w_i)\) is equal to 1, 2, or \(\geq 3\) \[49\]. Using these approximations, we set \(\hat{d}\) equal to \(d_1, d_2,\) or \(d_3\) respectively:

\[
\begin{align*}
    d_1 &= 1 - \frac{2n_2}{n_1 + 2n_2} \\
    d_2 &= 2 - \frac{3n_3n_1}{n_2(n_1 + 2n_2)} \\
    d_3 &= 3 - \frac{4n_4n_1}{n_3(n_1 + 2n_2)}
\end{align*}
\]

where \(n_i\) is the number of bigrams with frequency \(i\).

Using the language model for each sub-region, we estimate the probability that the social media posts emanating from it contain a query string \(q = (w_1, \ldots w_k)\) that represents a perception of interest. Based on these absolute probabilities across all sub-regions, we then compute the relative probability that the perception represented by query \(q\) is exhibited within sub-region \(\ell_i\) using Bayes rule:

\[
P(\ell_i | q) = \frac{P(q|\ell_i)P(\ell_i)}{\sum_j P(q|\ell_j)P(\ell_j)}
\]

\(P(\ell_i)\) is the prior probability that a social media post is from sub-region \(\ell_i\) and is given by \(N(\ell_i)/N(\mathcal{L})\). \(N(\ell_i)\) is the number of posts in \(\ell_i\) and \(N(\mathcal{L})\) is the total number of posts in the entire city \(\mathcal{L}\). Then, we define \(P(q|\ell_i)\) as:

\[
P(q|\ell_i) = \prod_{j=2}^k P(\ell_i, (w_{j-1}, w_j))
\]

### 2.2. Model Enhancement

Equation 8 computes the probability of a perception within a sub-region only based on the language model trained using social media posts within the sub-region. Previous works, however, suggest that the proximity of the users to an object or an event increases their propensity to post about it \[15, 46\]. In other words, the unigrams and bigrams representing the language of a sub-region may be influenced by the landmarks and events within its neighboring sub-regions that people discuss.
on social media. To model this influence, rather than directly using $P(q|\ell_i)$ to estimate the probability that a query $q$ arises from sub-region $\ell_i$, we instead apply a geo-smoothing function $\Theta^\circ(\ell_i|T)$ that combines the posterior probabilities of the neighboring sub-regions to consider their influence on the language in $\ell_i$.

The most important consideration in the design of $\Theta^\circ(\ell_i|T)$ is the form of a decay component that reduces the contribution of neighboring regions as they lie increasingly farther away from $\ell_i$ [33]. Comparative studies confirm that geo-smoothing performs best when this decay component takes a polynomial form [15, 46]; we thus consider a simple polynomial term that has been effective in geo-locating documents [46]. Letting $\Omega_k(\ell_i)$ be the set of neighbors of $\ell_i$ whose distance is $k$ cells away, and $P_{\ell_i}(T) = P(\ell_i|T)$, we define this geo-smoothing function as:

$$\Theta^\circ(P_{\ell_i}(T); \alpha, d) = (1 - \alpha)P_{\ell_i}(T) + \alpha \sum_{k=1}^{d} \sum_{\omega \in \Omega_k(\ell_i)} \frac{P_\omega(T)}{(2k + 1)^2 - 1} \quad (9)$$

where $\alpha \in [0, 1]$ is a smoothing weight specifying how strongly neighboring probabilities should be considered, and $d$ is the smoothing diameter, that is, the largest distance from which a neighbor can be located.

The choice of the hyperparameters $\alpha$ and $d$ control the number and the extent to which the neighboring sub-regions are factored into the computation of the likelihoods. Because of a lack of a priori information, we set $\alpha = 0.5$ so that the effect of the language models of all the neighboring sub-regions and the local language model is considered equally. The value of $d$ presents another tradeoff. We would like $d$ to be as small as possible so that we do not add “noisy” information into the local language model based on irrelevant terms that appear in the sub-regions that are further away. However, we also want $d$ to be large enough to include all neighbors that may contain valuable, relevant information that should indeed be incorporated in $\Theta^\circ$. Therefore, we study the sensitivity of of $\Theta^\circ(P_{\ell_i}(T); \alpha, d)$ against the value of $d$ experimentally.

For this study, we assigned a random probability to $k = 100$ sub-regions such that $\sum_{i=1}^{k} P_{\ell_i}(T) = 1$, and then compute $\Theta^\circ(P_{\ell_i}(T); \alpha, d)$ at $\alpha = 0.5$ for increasing values of $d$. We repeated this experiment a large number number of times with different random assignments of $P_{\ell_i}(T)$ to understand how these probabilities may impact $\Theta^\circ$ for different $k$. Figure 1(a) shows the change in $\Theta^\circ$ for every $\ell_i$ as a function of $d$ for 1,000 experimental runs. Each thin blue trend corresponds to a single run of the experiment. We also plotted the average value of $\Theta^\circ$ for every $k$ across all runs as a red, thick line and the 95% confidence interval bounds on this mean as the green bounding trends. We observe a gradual convergence of $\Theta^\circ$ across all runs irrespective of the initial assignments of the probabilities to the sub-regions, and the 95% confidence interval of the mean slowly widens with increasing $d$. Figure 1(b) plots the difference in this mean value as a function of increasing $d$, demonstrating an exponential decay in the average change of $\Theta^\circ$ as $d$ rises. The shoulder of this decay, that is, the point at which the change in $\Theta^\circ$ begins to decrease
significantly, falls at approximately \( d = 5 \). Based on: (i) the gradual convergence of \( \Theta^\circ \) with increasing \( d \); and (ii) a significant reduction in the average change of \( \Theta^\circ \) at \( d = 5 \), we decided to use this value in our illustrations.

2.3. Model Visualization

The concentration of a perception \( q \) within a sub-region is given by the relative probability that the language model for that sub-region generates \( q \). Depending on the perception and reasons behind a query, it may be appropriate to visualize these relative probabilities using a discrete or a continuous representation. In the discrete representation, we represent the concentration of a perception within each individual sub-region as is, or using the actual values. In the continuous representation, we fit a non-parametric density \( \hat{f}_H(x) \) across the relative probabilities across all sub-regions to smoothly model the regional spread of the perception. This non-parametric density takes the form [50]:

\[
\hat{f}_H(x) = \frac{1}{n} \sum_{i=1}^{n} K_H(x - x_i) 
\]

with

\[
K_H(x) = |H|^{1/2} \Phi(H^{-1/2}x)
\]

and \( \Phi \) as a multivariate kernel function that defines the shape of the density estimate. \( H = \text{diag}(h_i) \) is a positive definite 2 × 2 bandwidth matrix that controls how the length and width of the kernel function may be affected by the probabilities. Without prior information about how the likelihoods of perceptions may change as we move farther away from the sub-regions with a very high probability, we use a Gaussian kernel function that is symmetric about its center and exhibits a smooth
We define the components of $H$ using a well-established heuristic [47] for choosing the bandwidth matrix with a multivariate Gaussian kernel:

$$h_i = 1.06 \min(\sigma_i, R_i/1.34)n^{-1/5}$$  \hspace{1cm} (13)$$

where $\sigma_i$ is the standard deviation and $R_i$ is the range of the interquartile values distributed along dimension $i$ (latitude or longitude). We note that in practice perceptions may not decrease in a smooth, continuous fashion as we move away from high probability sub-regions. However, the continuous representation allows us to estimate the “reach” or spread of a perception broadly, based on the positions of the sub-regions with high probabilities.

Depending on the type of the query and the intent of the analysis, either the discrete or the continuous representation may be more suitable. Because the continuous representation allows us to view broad diffusion or spread, it is suitable for understanding perceptions that are relevant to the entire region. In contrast, the discrete representation identifies specific “pockets”, and hence, is more suitable for perceptions that are of interest to specific positions and sub-regions. To illustrate the choice of discrete or continuous representation, we consider concentrations of the perception “traffic alert” across a specific region in Figure 2. Usually, traffic alerts are shared only from specific positions, and not pervasively through the region. We should thus analyze this perception using discrete visualization in Figure 2(a) to 

\[ \Phi(x) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}x^T x\right) \]  \hspace{1cm} (12)
find individual sub-regions from where these alerts are originating, including major bridges and U.S. Routes and Interstates. In contrast, the continuous representation of the same perception in Figure 2(b) does not distinguish between these positions. In fact, because traffic alerts are patchy through the region, the fitted density extends across many sub-regions where there are no specific reported instances. Thus, in this case, using the continuous representation obfuscates our ability to identify those sub-regions where traffic alerts are really relevant. On the other hand, Figure 3 shows the results of querying “traffic” across another region, labeled Region #2, that is more uniform and densely populated. Figure 3(a) shows that the perception of traffic is strong throughout this second region, with a higher concentration in the center of its lower and upper parts. The continuous representation in Figure 3(b) better conveys this widespread perception, and highlights the radius of its two foci. Thus, we use the discrete representation for perceptions that pertain only to specific sub-regions, and turn to the continuous representation for perceptions that may be spread regionally.

2.4. Model Application

Figure 4 depicts the workflow process used to apply our probabilistic language modeling approach to extract perceptions from a given geographic region. These steps are as follows:
(1) Tweets shared in real-time within a geographic region are collected through Twitter’s streaming API\textsuperscript{4}. The locations parameter with the API function call for retrieving tweets specifies a bounding box within which all collected tweets must originate. Moreover, only those tweets that have embedded geo-coordinates are collected. Previous studies confirm that with the “locations” parameter, the API will return every tweet with geo-coordinates, whereas other parameter settings return just a small random sample of all the real-time tweets \cite{42}.

(2) The next step is to divide the collection of tweets into sub-regions according to their geo-coordinates. The number of sub-regions selected for partitioning should be according to the resolution of the results desired. Dividing a region into a large number of sub-regions can pinpoint perceptions more precisely, however, this requires sufficiently high tweet density to accurately train the language model within each sub-region. If the tweet density across the region is too low, or if the distribution of tweets is non-uniform causing some sub-regions to contain a small number of posts, the high resolution results may be generated by inaccurate language models that have not been trained over enough information.

(3) The third step is to build geo-smoothed language models $\Theta^o(P_t(T); \alpha, d)$ for each sub-region using the collected data. To build these models, the researcher

\textsuperscript{4}https://dev.twitter.com/docs/api/streaming
must choose appropriate values of \( \alpha \) and \( d \) based on the knowledge about how his perceptions of interest may be impacted by regional characteristics. For example, the perceptions about a local event may be modulated by the proximity to the event, making the value of \( \alpha \) close to one appropriate. Without any \textit{a priori} knowledge, one can use \( \alpha = 0.5 \) to equally weigh the language models of a sub-region and its neighbors. The best value of \( d \) can subsequently be found through a sensitivity experiment such as the one described in Section 2.2.

(4) In the last step, a sequence of terms \( q \) is used to query the ensemble of language models. The relative probabilities of the queried perceptions within sub-regions can be visualized using both discrete and continuous visualizations.

We note that the language models can be updated at regular intervals as additional real-time tweets are collected. Depending on the application, the language models may be trained using only a recent set of tweets (to study changes in perceptions over time) or may be continually updated with both past and present tweets (to identify long-standing perceptions across a geographic region). We have made available Python and R code for running this entire process online at \url{http://www.cse.uconn.edu/~ssg/twitter_perceptions.zip}.

3. Data Collection and Preprocessing

We seek to illustrate the application of our perception mining approach for discovering and monitoring road traffic and mass transit conditions in cities and suburban areas. Issues related to transportation have now assumed serious proportions worldwide, thanks to our society’s reliance on motorization and inflated population densities around urban hubs [5], making it a very timely theme for illustrating our approach. It is believed that road traffic congestion degrades the efficiency of the transportation infrastructure and can have a severe economic impact on businesses and organizations [5]. Traffic congestion is also intimately related to the quality, frequency, and reach of mass transit; infrequent and tardy mass transit leave people with few options but to use the road network. In order to devise and manage a combination of road and mass transit networks that can efficiently mitigate congestion [4], it is essential to understand the seasonal and non-seasonal patterns in their usage through continuous data collection and monitoring. Because traffic delays are extraordinarily frustrating for occasional travelers and commuters [38, 35, 27] alike, we postulate that they frequently turn to social media as a means to vent their problems, observations, and thoughts as they face a variety of conditions. Our preliminary work confirms such use of social media to report on the ground conditions [19]. Thus, mining perceptions across the transportation infrastructure may reveal important details about where, why, and how various traffic events and conditions emerge.

To illustrate the use of perception mining in transportation planning and control, we harvested a large collection of geo-tagged feeds from Twitter between June 11\textsuperscript{th} and November 26\textsuperscript{th}, 2013 across four different regions, namely: (i) Midtown
and downtown Manhattan in New York City (NYC), (ii) Washington DC (DC); (iii) the greater Kansas City (KC) area; and (iv) the state of Connecticut (CT). The approximate number of tweets from each region are listed in Table 1. These four regions were chosen because they have distinct regional and social characteristics and tweet densities, however, all the regions include major cities whose traffic infrastructure is routinely strained. For example, the NYC region is the densest in terms of both population and social media usage, with an incredible 111,437 tweets per square kilometer. Washington DC is also a major metropolitan area but shows a decidedly smaller tweet density at 18,610 tweets per square kilometer. The KC region is centered around Kansas City, Missouri, which is another major city. However, compared to NYC and DC, the KC region is much broader and encompasses a vast suburban sprawl including towns and roadways that lie beyond the city limits. The tweet density of the KC region is 1,557 per square kilometer, which is an order of magnitude lower than the density of the DC region. Finally, the CT region covers an entire state and includes pockets of large cities and population hubs, but is mostly composed of rural farmlands, forests, and coastal regions. With a tweet density of only 318 tweets per square kilometer, the use of Twitter is even less prevalent across CT compared to the KC region.

To build our language models, we divided each region using a 15 × 15 lattice that yields 225 non-overlapping sub-regions; each with the same area. This uniform partitioning allows us to demonstrate how our approach can identify perceptions at varying levels of granularity depending on the size of the original region. For example, given the small size and high tweet density of NYC, the language models may identify perceptions to within a couple of kilometers. Thus, “block level” NYC observations may reveal issues such as availability of parking and transportation, and local, street-wise traffic patterns. Although the DC and KC regions have lower tweet densities because they represent broader regions, perceptions will affect entire districts or at the “district level”. Such aggregate observations include spots with tendencies for traffic congestion, and issues affecting daily commutes into and out of the cities. Finally, although the CT region is larger and has an overall lower tweet density, much of the tweeting activity is focused around major cities and interestingly, on the major interstates that cut through the state as illustrated in Figure 5. “State-level” perceptions in CT may thus identify the most problematic areas for traffic congestion and underlying problems throughout the state, providing essential information to its department of transportation.

We pre-processed the tweets from every region by eliminating non-english words, hashtags (words starting with “#”), and usernames (words starting with “@”). We also transformed all the words to lowercase and stripped away all punctuation and links to Web pages. We also eliminated the 200 most frequently used words from the tweets in each region. This set typically contained words such as “at”, “the”, and “or”, which lack contextual information, and hence, introduce noise in the estimation of useful bigrams. The size of our stopword list was chosen to be ap-
approximately equal to 0.5% of the number of distinct words across NYC, which is the largest corpus across all four regions. We also include a “catch-all” unigram “<misc>” to aggregate the probability of words that occur only once. This term thus accounts for the many miscellaneous, shorthand, mis-spelled, and other user-specific notations that are common on Twitter. Across the four data sets, only a small percentage (2-3%) of words were mapped to this catch-all unigram, suggesting that we can control this source of distortion without impacting the fidelity of the model.
Table 1. Region-wise Summary of Tweets

<table>
<thead>
<tr>
<th>Region Size</th>
<th>NYC</th>
<th>DC</th>
<th>KC</th>
<th>CT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region Type</td>
<td>Metro City</td>
<td>Metro City</td>
<td>City Sprawl</td>
<td>Urban State</td>
</tr>
<tr>
<td>Sub-region Type</td>
<td>Block-wide</td>
<td>District-wide</td>
<td>District-wide</td>
<td>State-wide</td>
</tr>
<tr>
<td>Tweets</td>
<td>5,726,769</td>
<td>3,011,841</td>
<td>4,297,746</td>
<td>3,028,275</td>
</tr>
<tr>
<td>Tweet Density</td>
<td>111,437/km²</td>
<td>18,610/km²</td>
<td>1,557/km²</td>
<td>318/km²</td>
</tr>
</tbody>
</table>

4. Illustrations

In this section, we extract perceptions from Twitter feeds across NYC, DC, KC, and CT regions. Within the transportation domain, we focus our illustrations on understanding the effects of construction zones and road work, rush hour commutes, and service delays in mass transit systems. Ideally, the concentration of perceptions within sub-regions revealed by our probabilistic approach must be verified quantitatively and compared against a series of baseline methods. However, because a “perception” is an emotional, qualitative concept that may be modulated by countless feelings, beliefs, and observations of individuals, a quantitative comparison of the computed probabilities with actual perceptions is infeasible. Furthermore, limited efforts that analyze social media to understand an environment are geared for specific purposes and information [44, 14, 54]. We view our model as being a source of emotionally driven data that is complementary, instead of directly competing, with these prior efforts.

Since quantitative validation of the results produced by our model is all but impossible, in our illustrations, we confirm that our perceptions represent on-the-ground conditions using publicly available data sources that include online databases, local news stories, and government reports. As each illustration shows, we find clear associations between the quantitative probabilities extracted from our model and the information obtained from these sources.

4.1. Construction and Lane Closures

Many studies have shown that street and highway lane closures due to road work in construction zones are among the leading causes of traffic congestion [43]. To predict and minimize the impact of delays caused by these road improvements that are often necessary, transportation agencies employ numerous kinds of traffic models [9, 1, 24], data collection technologies [28], and practical reports about successful road construction projects [25, 32]. Despite the use of these tools, construction zones and lane closures cause delays and disruptions that are commonly longer and farther reaching than anticipated [41]. Monitoring ground truth traffic conditions as experienced first hand by the drivers driving through these work zones, can allow transportation agencies to make reactive, short-term adjustments to the pace and intensity of the road work so as to minimize its impact. In this section, we illus-
trate how the positional likelihoods extracted from social media posts may provide information for making such adjustments. Given our interest in understanding the widespread effects of construction work on traffic, we visualize these perceptions using the continuous representation.

Figure 6 visualizes the query “construction zone” across the DC and NYC regions. As the figure illustrates, we find construction to be prevalent across both major metropolitan cities, particularly in downtown Washington DC. Figure 7 visualizes the same query across the CT and KC regions. We find perceptions of lane closures and construction zones in Figures 7(a) around I-95 by Bridgeport and New Haven, I-91 near New Britain, and on I-91 and I-84 by the city of Hartford. Schedules of road work at these cities during our data collection period confirm that these perceptions refer to road construction projects at these locations. Moreover, the locations and spread of these perceptions suggest the wider extent to which drivers are impacted by these construction projects. For example, we find that additional road work that occurs throughout Hartford may have attributed to perceptions that extend across a continuous 18 km stretch.

Figure 7(b) shows an expanse of perceptions revealed by the query “construction zone” across DC and NYC.
zone” in the KC region centered on U.S. Routes 71 and 50 in Grandview, Missouri, which can attributed to the construction of a new bridge deck for the interstate system that impacts both highway and local traffic. Agencies may find multiple uses of this information while planning future construction projects within these areas. For example, they may inform Grandview citizens, and take measures to alleviate and reroute traffic in the residential area. Similarly, the state of Connecticut may note the need to alert drivers about upcoming construction projects in a wider area around Hartford and New Haven than originally anticipated. Moreover, they may also wish to suggest alternate routes to reduce the number of drivers passing through these zones.

4.2. Commuter Rush Hour

Commuters who travel to and from work during the same “rush hour” represent a major source of strain on the transportation infrastructure [20]. In this section, we examine perceptions on rush hour traffic through the KC and CT regions, home to major cities with suburban sprawls from where people commute daily. Because commuter traffic and the associated perceptions spread through a city rather than being localized to specific streets or intersections, we use the continuous representation to visualize these perceptions.

Figure 8 shows that the perception “traffic during commute” is spread through downtown Washington DC and across Manhattan. We find large pockets in NYC adjacent to the Brooklyn Bridge and Lincoln Tunnel; these two represent primary

\[\text{bit.ly/I1ND8f8}\]
conduits out of the city to Brooklyn and New Jersey respectively; which are homes to many commuters. The DC region has an intense concentration of perceptions in the downtown district where most federal agencies and government offices are located. Another stretch of this perception can be found around University Heights and Brightwood Park, neighborhoods where many city dwellers live, as well as around K and H street NE, which are thoroughfares known to have the offices of many political think tanks and lobbyists\(^h\).

We next turn our attention to commuter traffic going in and out of major cities from suburbs in the CT and KC regions in Figure 9. Figure 9(a) shows widespread perceptions of commuter traffic across the state of CT. In particular, we observe very strong perceptions centered around New Haven and Hartford; which is not surprising as these cities are ranked as having the 19\(^{th}\) and 30\(^{th}\) slowest Friday rush hour commutes in the United States\(^i\). These perceptions are highly concentrated and do not extend beyond one or two towns away from the city of New Haven. On the other hand, perceptions around Hartford spread far beyond the state’s capitol, covering an approximate area of 877 km\(^2\) from Bristol to Rockville. This high-level perspective may suggest to the state officials that the impact of Hartford rush hour

\(^h\)wapo.st/1btq8X4
\(^i\)bit.ly/1lRU21C
traffic is so wide and significant that it may actually require more attention than the nationally recognized rush hour traffic problems that surround New Haven.

Figure 9(b) visualizes commuter perceptions surrounding Kansas City. Although Kansas City is small compared to many other cities in the U.S., it faces major commuter problems and ranks in the top 100 cities with the slowest Friday commute to work. The figure identifies large concentrations of commuter traffic perceptions within the city and near a three-way intersection of Interstates 35 and 70, and U.S. Route 71. The model also reveals pockets of commuter traffic perceptions within the suburban sprawl around the city. Some of these pockets correspond to local conditions such as construction that were identified in Section 4.1. Other pockets, such as the ones located in Olathe and Overland Park, Kansas, span residential streets and communities that may be congested due to single lane commuter roads. Additionally, these perceptions may also be shared by residents who are about to begin or have just finished their commute, and are expressing thoughts about the traffic they face(d). The multitude of pockets identified correlates with a 2006 report from the U.S. Census Bureau that identifies Kansas City as having a low commuter efficiency score, which they define as the number of commuters who carpool or take mass transit divided by all commuters [6]. Operating at just 12% efficiency (large metropolitan cities like Boston and Los Angeles, by comparison, operate at 23% and 20%), most commuters travel alone and in their own vehicles, adding to overall congestion. To decrease congestion and improve the perception of commuting difficulty, Kansas City officials, in the light of this information, may decide to create and promote ride share or carpooling incentive programs. Thus, in case of both the CT and KC regions, our mining highlights roadways with the strongest perceptions of rush hour traffic, giving state and city agencies quick information about the most stressful and difficult roads that travelers use on a daily basis.

4.3. Mass Transit

Both New York City and Washington DC feature a massive public transportation infrastructure that commuters and citizens use on a daily basis; the Metro Transit Authority (MTA) of New York City operates and manages a system that over 5.4 million people ride annually\(^1\) while the Washington DC Metro system services more than 750,000 riders daily\(^2\). Given the magnitude of these mass transit systems, delays are a norm for a number of reasons including overcrowded buses and subways, slow service at major stations, as well as a host of other unpredictable and unanticipated problems. Additional factors causing delays may include road or track segments with speed restrictions, increased foot traffic at some stations, and proximity to businesses and corporations where all employees commute at similar times. Given their sheer volume of riders, it is nearly impossible for these transit

\(^1\)http://web.mta.info/mta/network.htm
agencies to survey for qualitative data about where and how their riders perceive extensive delays and “problem spots” in the infrastructure. Thus, we demonstrate how these agencies may instead turn to our language model to explore perceptions of bus and subway delays, to collect such data quantitatively, quickly, and at no cost. We utilize discrete visualizations to identify the specific pockets and locations where thoughts and observations about service delays are the strongest.

Figure 10 shows perceptions of bus and subway delays across Manhattan. We find such perceptions to be concentrated in midtown Manhattan in sub-regions that host Grand Central Station, Penn Station, and the Port Authority bus terminal, which are the busiest commuter rail and bus stations in the city. Both discrete visualizations identify strong perceptions around West 14th St and 8th Ave., which contains subway stations for the A, C, E, and L lines. This pocket of delays corresponds to the scheduled work on these lines during our data collection period. During these times, the agency provided free shuttle and bus service since trains were unable to pass through these stations. Other disruptions to service, include a shutdown of the R line that brought more passengers to the C line and water main breaks that led to service changes. These have collectively resulted in long-standing perceptions that both subway and bus service is tardy in this sub-region. This analysis shows the severely disproportional impact of service disruptions at the stations in these sub-regions, and how these disruptions affect the riders’ perceptions of the MTA.

\[2\] bit.ly/1eOnEft
\[3\] bit.ly/1f7VYRt
\[4\] cbsloc.al/1fKAYEO
Compared to NYC, perceptions about delayed mass transit across DC as visualized in Figure 11 show up in fewer sub-regions. The strongest perceptions of late bus service can be found next to Nationals Park, which is a baseball stadium where the district’s major league baseball team plays their home games. Eight different
bus lines make stops at the stadium, to accommodate the surge of riders during home games\(^6\). However, traffic around the stadium is constantly snarled following a game\(^7\), which inevitably causes long delays. Spectators at the stadium may even be stranded because of service delays on the bus routes that extend beyond the time when bus and subway service shuts down, turning to Twitter to voice their frustrations and concerns\(^8\).

Figure 11(b) shows strong perceptions of late subways within downtown sub-regions that lead to the Capitol Building, the White House, and the National Mall. The strongest perceptions surround the Metro Center station, which is the largest and most central subway station across the entire system. Because of this central position, Metro Station patrons are frequently impacted by service disruptions and delays stemming from other stations across all lines, including a train that was derailed during the data collection period\(^9\). In conclusion, our perception mining approach was able to pinpoint two mass transit stations that riders most perceive as having extensive delays. The mass transit agency can use this knowledge to either improve the impact of service disruptions at these stations, or in reality if the delays are not severe, undertake promotional campaigns to improve riders’ perceptions about them.

5. Related Research

In this section, we compare and contrast related work along two dimensions, namely: (i) analyzing social media posts; and (ii) monitoring transportation landscapes.

Different modeling approaches interpret and extract intelligence from social media posts. They differ according to the incorporated features, which include spatial word usage [15,13], posting behavior [37], check-in sequences in one’s social network [45], and latent topics [29]. This intelligence has been utilized for applications such as point-of-interest extraction [34], city discrimination [30], attaching Wikipedia articles to geo-positions [51,8], and sentiment analysis [2]. Similar to our work, this intelligence has also been used to discover details about a geographic area. For example, Purohit et al. developed a platform that aggregates and analyzes various dimensions of citizens reports about an event including the specific topics and their associated sentiment [44]. Gao et al. propose a system that extracts information from social media to coordinate the delivery of aid to crisis-stricken areas [23]. Cranshaw et al. use spectral clustering to identify local neighborhoods in an urban area based on whether similar types of users check into the same locations [14]. Yuan et al. use multiple social network and social media accounts in a computational framework that can categorize different types of urban lifestyles [54]. Note that these previous works concentrate on discovering a specific aspect or qual-

\(^6\)http://www.wmata.com/getting_around/metro_events/nationals/
\(^7\)wapo.st/1f5yED
\(^8\)bit.ly/1bzE7uS
ity about (typically) urban environments that have high tweet density and where most citizens are expected to be active social media participants. In this work, we tie positional information of social media posts to the thoughts, feelings, and actions that citizens within any environment convey. We present a generic model that pinpoints the degree to which people *en masse* exhibit a perception. Our approach is thus complementary to the proposed methods for extracting specific aspects from social media. Thanks to its generic capabilities, it can be deployed with any of the contemporary approaches.

Because our illustrations are anchored around extracting traffic and public transportation conditions, we also compare our work to the many innovative approaches proposed to monitor such conditions. These include detecting traffic parameters using computer vision algorithms on camera feeds [22], through simulations [48], employing a geo-processing traffic monitoring system [40], or by capturing measurements with RFID and GSM technologies [39]. Although model- and simulation-based approaches may be able to accurately predict future conditions based on normal and regular traffic patterns, they cannot anticipate delays that occur due to unexpected events such as emergency construction, traffic accidents, or weather conditions. Such unexpected events could be detected from real-time measurements using embedded sensors. However, this requires sensor devices to be deployed on the road or in vehicles, which may be very costly and unaffordable for many cities.

In this paper, we used numerous illustrations to demonstrate how transportation and government agencies can turn to social media as new kinds of road sensors that can measure traffic and public transportation conditions at a moment’s notice and virtually for free. These social media sensors offer a unique type of data, rooted in peoples’ emotions and feelings, that can serve as an alternative to embedded sensors or even as complementary data sources. For example, perception mining can augment the quantitative readings provided by road sensors with information about the direct impact of the conditions on citizens.

6. Conclusions and Future Directions

In this paper, we introduced a new methodology for extracting the perceptions that people express within social media posts. These perceptions, which represent people’s thoughts, opinions, actions, and responses about the world around them, are mined using a probabilistic model that may be more suitable for social media posts as compared to mining them via semantic analysis. Our probabilistic approach uses an ensemble of language models that interpolate unigrams and bigrams, and are also smoothed across neighboring sub-regions to account for the influence of language and landmarks from wider areas. Using discrete and continuous visualizations, we illustrate the utility of our language model to identify problems encountered in the transportation domain. Specifically, we mine people’s reactions to lane closures, commuter traffic, and mass transit delays. We verify these perceptions using publicly available data sources and articles.
Our future work will concentrate on incorporating term disambiguation algorithms, lexical databases, and domain-specific ontologies into our approach. We will also research ways to improve the accuracy of the locations where a given perception is concentrated. Finally, we will study how people’s perceptions may evolve spatially and temporally.

References


[34] W. Li, P. Serdyukov, A. P. de Vries, C. Eickhoff, and M. Larson. The Where in the


