InMacs: Interactive modeling and comparison of sentiments from sequence data

Subhajit Das
das@gatech.edu
Georgia Institute of Technology USA
Atlanta, GA

Florina Dutt
florina.design@gatech.edu
Georgia Institute of Technology USA
Atlanta, GA

ABSTRACT

Machine learning (ML) has changed various problem domains by offering insightful solutions. For example, urban planners (usually non-experts in ML) model sequence data such as text using AutoML systems (e.g., OrangeML, Google Cloud ML, etc.). Specifically, these users mine unstructured text data using Twitter API to compare peoples' sentiment/opinion on urban spaces. However, the current AutoML tools restrict the active participation of end-users in model construction/adjustment. To resolve this problem, we designed an effective technique that combines an interactive visual interface with an AutoML model solver incorporating users' domain knowledge as feedback that adjusts the underlying models' behavior. In this paper, we present InMacs, an innovative visual analytics (VA) system that allows urban planners to interactively construct sentiment classifiers and visualize the output of these models to compare peoples' sentiment across multiple geolocations. Through a case study we discuss our on-going work with urban planners that includes design, build, and validation of our prototype. Furthermore, we discuss the effectiveness and the generalizability of our interactive technique on other domains by presenting a case study that compares business reviews from the publicly available Yelp dataset.

KEYWORDS

Sentiment classification, Topic Models, Sequence data, Twitter API, AutoML, Interactive machine learning, Domain experts

ACM Reference Format:
Subhajit Das and Florina Dutt. 2020. InMacs: Interactive modeling and comparison of sentiments from sequence data . In *Proceedings of (DaSH@KDD).* ACM, New York, NY, USA, 8 pages.

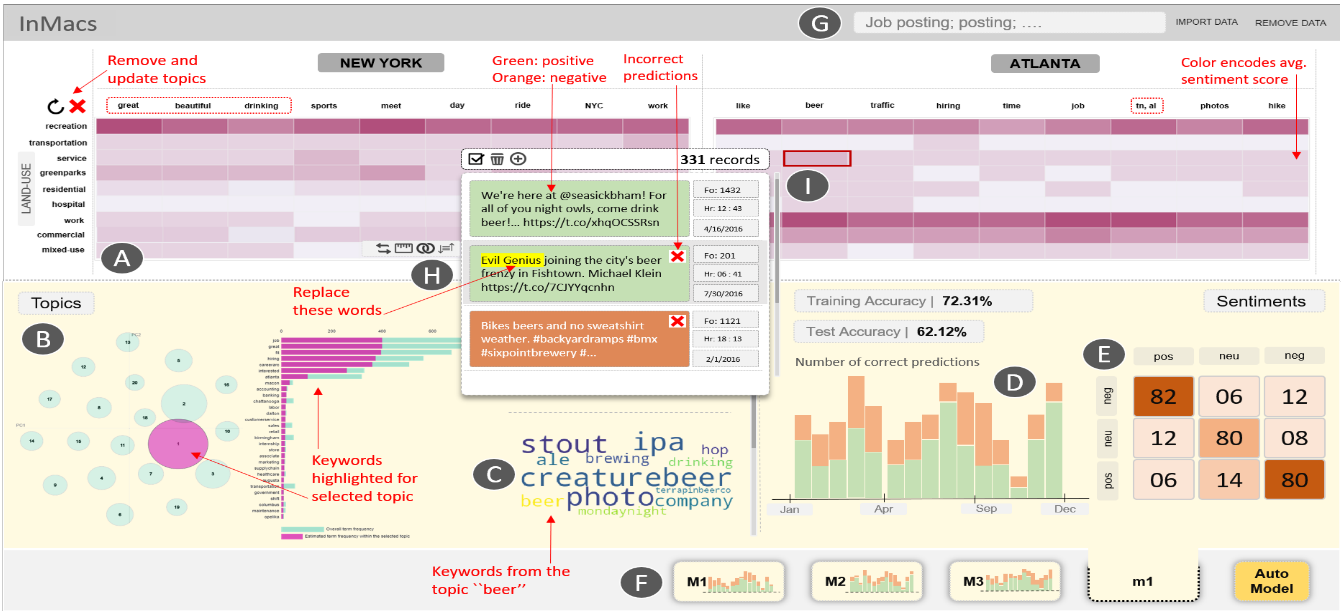
1 INTRODUCTION

Machine learning (ML) has been effectively used in many real-world data analytic problem scenarios (e.g., in marketing, finance, healthcare, etc. [54]). To further discover ML's real-world application, we worked closely with urban planners to learn that they deploy ML models to infer peoples' sentiments using unstructured text data. For example, using Twitter Short Text Posts (STP) [47] as a source, urban planners seek citizen's participation [13, 39] to know important nuances about places they frequently visit or live in. One of the critical tasks in their analysis process is to compare cities with varied urban characteristics to understand their differences/similarities in relation to domain-specific characteristics such as, walkability, safety, liveliness, etc. [5]. For example, a Tweet may report "I prefer walking along the river than being stuck in a car all day .....", highlighting peoples' preference over spatial qualities in urban spaces. Through repeated discussions with planners, we further learned that they sometimes work in collaboration with a data analyst who may have an intermediate understanding of ML. In this process they often learn applied ML coding skillsets (in Python or R programming environment) through tutorials or codes available online. However, most frequently they rely on AutoML systems to model text data satisfying their data analytic goals. For example, they specify a pre-annotated (often hand-labeled) dataset of Tweet corpus to an AutoML model solver (e.g., Auto-Weka, Orange ML, etc. [40]) to train a sentiment classifier. Next, based on the models' prediction, they compare peoples' sentiment (expressed in the Tweet posts) from different geo-locations.

Though current AutoML tools [34, 48, 58] are useful and effective in model construction, urban planners find them to: (1) inhibit incorporating their feedback as part of the models' training input, and (2) disallow including them as an active participant in the model construction/refinement process. There are also systems that either interactively construct/inspect sequence models [36] (e.g.,CNN, RNN, LDA etc. [12]) or compare social media data [30, 31, 60–62]. However, urban planners confirmed that there are none that allows user feedback to be integrated in an AutoML's pipeline as they explore and compare sequence data from different contexts (e.g., multiple cities, or other self-defined domain-specific categories). For example, the current AutoML pipeline [57] needs an input training data; based on which it generates multiple models and automatically selects the best model for the dataset with respect to a user-specified model performance metric (e.g., cross-validation score, precision, residual score etc.). However, if a user sees any discrepancy in the model, they cannot adjust its behavior by providing feedback to the modeling process or see changes in the model by interactively updating part of the input text data.

In this paper, we empower urban planners by designing an effective technique that combines an interactive visual interface with an AutoML model solver incorporating their domain knowledge as feedback to adjust underlying sequence models. As such, we present InMacs, a novel visual analytics (VA) system that allows urban planners to interactively construct sentiment classifiers and visualize the output of these models to compare peoples' sentiment from multiple geolocations (e.g., two cities) or from diverse topics of discussion. Past systems such as Matrix Wave [64] have looked at

DaSH@KDD, August 24, 2020, Virtual Conference



We're here at @seasickbham! For all of you night owls, come drink beer!... https://t.co/xhqOCSSRsn — Fo: 1432 — Hr: 12 : 43 — 4/16/2016

Evil Genius joining the city's beer frenzy in Fishtown. Michael Klein https://t.co/7CJYYqcnhn — Fo: 201 — Hr: 06 : 41 — 7/30/2016

Bikes beers and no sweatshirt weather. #backyardramps #bmx #sixpointbrewery #... — Fo: 1121 — Hr: 18 : 13 — 2/1/2016

	pos	neu	neg
neg	82	06	12
neu	12	80	08
pos	06	14	80

Figure 1: Main views of InMacs: A. Heatmap matrices. B. Topic scatterplot and bar chart view showing top keywords per topics. C. Word-cloud showing top keywords. D. Stacked bar chart showing number of correct sentiment predictions with time. E. Confusion matrix. F. Model overview showing sentiment classifiers constructed per iteration. G. Search bar to filter data. H. Text view to add/remove/replace and specify text length to training data. I. Mouse over heatmap to see Text View.

comparing event sequences on clickstream datasets and presented results as a matrix view. Similarly, Compare Cloud presented a VA system that helped users compare two media frames as wordclouds, grounded on text datasets [18]. While Matrix Wave serves clickstream data, Compare Cloud allows comparison of text data and is very close to what we are doing. Using these as inspirations, we design our visual platform that allows: (1) comparison of sequence data, and (2) interactive construction of sequence models, with user interface affordances that integrate domain expertise as feedback in the AutoML's model selection process.

Our technique comprises of two parts: (1) the frontend that creates interactive visualizations, and (2) the backend that implements the AutoML model solver injesting an input sequence data to construct a topic-context matrix M of dimension k by r (values of these can be interactively specified). Here k represents number of topics discussed in the text corpus, while r represents land use types (encoded using each Tweet's geolocation). These land use types are zones in the city such as *Work*, *Residential*, *Service*, *Transportation*, *Commercial*, *Mixed-use* etc. (can be retrieved using Tax Assessors dataset [3]). Thus the matrix presents the key topics discussed in the city through the lens of various land use types. Our technique visualizes the matrix M as a set of interactive heatmaps showing users an overview of the data. In InMacs, the heatmaps of multiple cities (see Figure 1-A) are visualised side by side adapting the “juxtaposition” technique to compare data as discussed by Gleicher et al.[25]. This helps urban planners understand how different cities compare in relation to topics discussed across all the land uses. Furthermore, users can interactively specify number of topics, change topic-keyword associations to adjust the construction of underlying topic models. Clicking on any cell of the matrix shows the set of Tweets of the said topic and land use. With this interaction users can find and remove noisy or outlier Tweets.

Further, we describe the results of our design process that is driven by incremental feedback from domain experts which led to multiple refinements of the tool. Finally through a case study we discuss the application of our technique with sequence data where urban planners model 200000 Tweets from the US Global dataset [2] comparing multiple cities. In addition we also present a use case with Yelp business review dataset [4], showing how InMacs can help a business strategist compare user reviews of various businesses (e.g, salons, cafes, bars, etc.) across different geolocations. This validates that our technique is effectively generalizable to other text datasets. In summary, we contribute the following:

- A visual analytic system to compare sequence data (text) using an automatic model selection (AutoML) approach.
- Findings of an on-going design study with urban planners leading to the development of a human in the loop based AutoML workflow.
- Two case studies that discusses the domain application and the generalizability of the presented tool.

2 RELATED WORK

From the literature we studied visual comparison of urban data, current AutoML tools, and interactive ML systems, as discussed below:
Visual Comparison of Data: Visual comparison of data has been widely used in various domains and problem cases such as comparison of website traffic flows, comparing the structure of protein sequences [29], comparing text data [30, 31, 60], etc. For example, MatrixWave facilitates visual comparison of event sequence traffic

patterns and allows users to interactively explore paths through websites [64]. SocialBrands is a VA tool designed to make sense of public perceptions on social media data and compare multi-dimensional brand personality of various brands using interactive visualizations [42]. There are also work that compares ML models to understand tradeoffs in model selection [19, 61, 62]. Law et al. [38] presented a taxonomy of pairwise comparison, based on which they presented Duo, a system that allows comparison of two sub-group of data items in a tabular dataset. Similarly, MS Excel also allows pivoting tables to compare sub-group of data. To understand visual comparison of data further, we studied the work by Gleicher et al. [25]. They discussed various visual comparison techniques such as juxtaposition, superposition, explicit encoding, etc. Other authors have used animation to visually compare information [9]. Inspired by these and other similar work [14, 27, 45] we used the juxtaposition technique to facilitate interactive comparison of urban data across multiple geolocation.

AutoML systems and pipelines: Recent years have seen a spur in the development of automated machine learning tools also called AutoML systems. These tools such as AutoWeka [35, 58], SigOpt [48], HyperOpt [10, 34], and many others [23, 40] seek to automate the process of model construction/selection enabling non-experts in ML to incorporate ML methods in their data analytic applications. For example, typically model construction requires specification of a problem case, input of training and test data, selection of a learning algorithm and associated hyperparameters. Next users need to find a ML library (e.g., Scikit Learn [51]) to construct/validate models. While AutoML automates this tediuous process, they fail to include user feedback into the modeling pipeline thus limiting the possibility of customization of model outputs that may better adhere to users' personal goals.

Holzinger et al. elicited that AutoML methods tend to be very useful in cases when there is easily available large static training data [28]. While that is true, in many cases the input data is noisy, contains error, or may need human intervention to validate them. For such cases, a new class of VA systems are being developed that looks at coupling a human with an AutoML model solver [1, 24, 57]. Snowcat facilitates inclusion of a human with an AutoML system [15] to perform a diverse set of ML tasks. Inspired by these systems, we seek to empower domain experts to adjust models by interactively cleaning/pre-processing the input training data such that AutoML model solvers may choose better performing models that more closely supports their goal.

Interactive model construction: While AutoML [11, 20] serves to readily solve the need to create models at the click of a button, interactive model construction incorporates human in the loop based dialogue between the user and the machine facilitating the possibility to externalize users domain knowledge into the models training process [21, 22]. This human-centered ML approach is further discussed by many researchers that see merit in including people in ML processes to account for human intent [6–8, 56]. Sacha et al. further discussed the application of VA systems as an interactive visual interface between automated algorithms and humans for effective data analysis [56]. This approach has previously been successfully applied to solve various ML tasks, including clustering, dimensionality reduction, regression, classification, etc. [16, 17, 19, 59]. Along the same lines, we seek to solve interactive modeling of sequence data to classify sentiments using human in the loop of current AutoML pipelines.

Visualizations and modeling in urban planning: Urban planners use large scale social media data [41, 43, 55] to get access to citizens' opinion [13, 39, 46] on topics related to their domain. Next they deploy various ML modeling techniques (e.g., topic models) and visualizations to make sense of the data [33]. For example, Zhang et al. discussed engaging citizens and other stakeholders in discussion related to spatial planning. In doing so, they demonstrated the application of a web-based toolkit applying hierarchical topic modeling [63]. Other approaches of topic modeling in urban planning using social media data can be seen here [26, 44]. Furthermore, the sentiment classification task has proven to be pivotal for urban planners to understand peoples' sentiment [33, 41, 53, 55]. For example, Paul et al. prototyped Compass, a deep learning based technique of spatio-temporal sentiment analysis from large-scale social media data, on the topic of US Election in 2016 [50]. While these works have been effective for urban planners, there are none in the literature that helps urban planners compare sequence data by using interactive visualizations. Based on our discussion with urban planners, we report that there are lack of visual interfaces that enable interactive construction of sentiment classifiers to compare social media data across multiple geolocations.

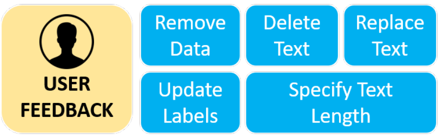


Figure 2: Types of user feedback to guide AutoML in InMacs.

3 DESIGN GUIDELINES AND TASKS

With repeated conversations with the urban planners, we framed the following set of tasks to design InMacs:

T1: Build topic models and analyse most frequently discussed topics in the text corpus. Find urban planning related topics from the set of retrieved topics.

T2: Interactively train sentiment classifiers to analyse sentiments and peoples' opinion across various urban planning related Tweets.

T3: Inspect broader implications of model output (topic and sentiment models) on input data in relation to multiple geolocations.

T4: Incrementally pre-process/clean training data to adjust the underlying models to suit their domain specific expectations (Figure 2).

T5: Compare peoples' perception on multiple geolocations with respect to their urban characteristics through the lens of most frequent urban planning related topics.

T6: Temporal analysis and comparison of text documents leading to a holistic understanding of peoples' sentiments and topics of discussion across many geolocations.

Motivated by these tasks we identified a set of design goals to address the data analytic goals of urban planners:

DG1: InMacs should allow visual comparison (with time) of urban characteristics and domain-specific topics in the data **(T1, T5, T6)**.

DG2: InMacs should empower users interactively construct sentiment classifiers, inspect sentiments across geographical locations and compare their temporal evolution **(T2, T6)**.

DG3: InMacs should facilitate the interactive adjustment/refinement of input training data to directly impact the creation of topic models and sentiment classifiers by the AutoML model solver **(T3, T4)**.

4 INMACS: USER INTERFACE

We present our system InMacs, that compares text data and interactively constructs sequence models containing the following views:
Heatmap view: This view visualizes a set of matrices (each representing an entity such as a geolocation) of *topics* and *categories* as a heatmap (see Figure 1-A). Users can interactively add, remove, and search new topics (**DG1**). They can click on any cell of the matrix to see the underlying text documents (e.g., Tweets) as a Text View (see Figure 1-I). They can hover mouse to see the input data linked with the Timeline view described below. The color of each cell may encode either the number of text documents per cell or the average sentiment score (normalized between 0-1 , 0: negative, 0.5: neutral, and 1: positive) for all the text documents in the cell. We decided to use a heatmap representation for the matrix to: (1) provide users an overview of the data, and (2) to compare sequence data side by side.
Timeline view: This view shows a histogram representing number of text documents on a time line (x-axis as the time axis). On the top half it visualises the data for one geolocation, while on the bottom it visualises the data for the other facilitating side by side comparison (**DG1**). Users can brush over the timeline to filter the heatmap to inspect text documents from a certain time/date range. If there are more than 2 entities or geolocations, then users can select the pair of geolocations to compare with (see Figure 4-B).
Text view: This view shows a list of text documents along with a set of other variables from the input data (e.g., number of followers, retweet count, date, etc.). More variables can be added on users request. The text is color encoded with its sentiment class (e.g., green for positive and red for negative sentiment). Incorrect predictions are shown with a red cross icon as seen in Figure 1-I, H. Furthermore, users can can remove, add/edit, or replace part of the text (words or n-grams) for multiple documents (**DG3**). In addition they can edit/update the target variable (text sentiment). By brushing their mouse on the text, users can specify the length of the input text for selected data instances.
Model overview panel: This is a horizontal shelf view visualizing the set of models (shown as thumbnails) constructed by AutoML (per iteration) as users interact with the data (see Figure 1-F). Each thumbnail gives a preview of the sentiment classifier showing a stacked histogram of correct/incorrect text documents on a time axis. The thumbnails can be clicked to see the details of the topic and sentiment classification model. The set of all topics are shown as a scatterplot, size of the circles encoding weights of the topics. A horizontal bar-chart shows the weights of the associated keywords per topic (see Figure 1-B). This view also contains the *Auto-Model* button that triggers the AutoML model solver to find an optimal topic and sentiment classification model (**DG2**). The sentiment classifier's output is visualized as a confusion matrix showing the training and test set's prediction accuracy. Each cell of the matrix is interactive; users can hover to know model accuracy details, click to update the heatmap view with data instances of the chosen class label. This view also shows a stacked histogram of correct (colored green) and incorrect (colored orange) text documents (Figure 1-D).

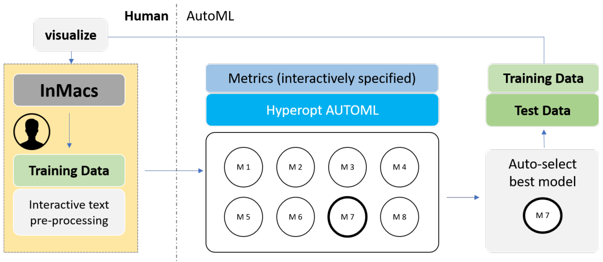


Figure 3: InMacs' design showing user input to AutoML.

5 TECHNIQUE

InMacs uses Hyperopt [10] as the Auto-ML model solver to search for an optimal topic model T and an optimal sentiment classifier C (see Figure 3). Users interact with InMacs to prepare the training data U for modeling and inspects the output of the models on both U and V (test data) to specify further inputs to the model creation process. Under the hood, Hyperopt is pre-specified with a learning algorithm (e.g., *CNN-Text model* [32]) and a set of hyperparameters (e.g. learning-rate). Per iteration, Hyperopt constructs n=200 models (n can be interactively adjusted). Hyperopt is also pre-specified with a metric score S that is utilised to select the optimal model T and C.
Pre-processing: Each data sample in the input data corpus is a list of words W, with data attributes such as *longitude*, *latitude*, *creation-date*, etc. W (retrieved from U and V) is pre-processed using standard text modeling approaches that includes, tokenisation, stemming, lemmatization, removal of stop words and special characters etc. [53]. Further pre-processing is applied as users interact and explore the data in InMacs (e.g., removal of text, specifying text length, replacing text, etc. as seen in Figure 2).
Topic Models: T constructs a topic-context matrix M of dimension k topics by r land use types (can be interactively specified). r is retrieved for each Tweet (W) based on their geolocation property. These land use types are zones in the city which may include *Work*, *Residential*, *Service*, *Transportation*, *Commercial*, *Mixed-use* etc. InMacs constructs M for each geolocation and then presents to users what key topics are discussed in relation to the land use type (e.g., 'job related topics in office districts represented by the *Work* land use). Each cell in M stores the *id*'s of the set of Tweets that was geolocated in the a specific land use type and contained text of a specific topic. Hyperopt is pre-specified to construct topic models (built by Scikit-Learn [51]) using learning algorithms LDA and NMF [12, 37]. Hyperopts adjusts these algorithms by tuning their hyperparameters such as *number of components*, *beta-loss*, *max-iteration*, *alpha*, etc. The pre-processed data W is vectorized as F to get the frequency of tokens. Using the Tfidf-Vectorizer module in Scikit-Learn, our technique constructs a Tfidf matrix TF, which is an input to Hyperopt. Next, the AutoML model solver constructs a set of sentiment classifiers C to retrieve sentiment class labels (e.g., *positive*, *negative*, and *neutral*).
Sentiment Classifiers: Our technique uses a CNN model C for sentence classification as described by Kim et al. [32]. To tune C, Hyperopt varies a set of pre-specified hyperparameters such as *learning-rate*, *number of convolution layers*, *drop-out-rate* etc. We used the Torch-Text Python module to represent the data as a tensor of indices based on a vocabulary. Furthermore, to represent each

input text data as a l x d matrix we used word embeddings from the Globe word embedding library [52] (l is length of a sentence, d is the word embedding dimension). Our technique constructs the sentiment classifiers using the Pytorch module [49] for Python 3.6.
User Feedback: Users can adjust W from the training data U with m data samples as they continue analysis in InMacs. For example, they may inspect the data to find q outlier or noisy data samples, thus updating $U = U_m - U_q$. Furthermore, they may trim the length of the input text to l (max. number of tokens) for a subset of data instances $U_s = u_1, u_2, ..u_s$, ($U_s \subseteq U$). To reduce ambuiguity in the models' reasoning on mis-leading input text tokens, users can also replace, delete, add words or n-grams to a subset of the training set U_p, where $U_p \subseteq U$. Furthermore, they can update, or correct ground truth labels G of the input training data U. These operations (see Figure 2) directly adjusts the underlying models' learning as it injests users domain knowledge in the training data.

6 CASE STUDY 1: US TWITTER DATA

Consider Trace is an urban planner who intends to use a text corpus of two hundred thousand Tweets [2] to mine and compare peoples' sentiment between New York City and Atlanta , USA. The data contains columns such as *created-at*, *tweet-text*, *followers*, *retweet-count*, *longitude*, *latitude*, etc. For each Tweet, given a pair of *longitude* and *latitude*, a *land use* is queried from the *Tax assessor's dataset* [3]. This adds a new column to the dataset representing the respective land use of the city such as *commercial*, *residential*, *mixed-use*, *institutional* for a given Tweet location. The data is annotated with sentiment class labels: *Positive*, *Negative*, and *Neutral.* They split the data into a training set of one hundred fifty thousand samples and the rest as test samples to construct sequence models. Trace loads the data in InMacs to continue their analysis.

For both cities InMacs builds a topic model and shows top eight topics. Trace sees a pair of heatmap views where every cell in the matrix shows Tweets of a certain topic occurring in a certain land use category (see Figure 1-A). Furthermore, the color encodes the number of Tweets per topic and land use. Next based on the color encoding of each cell, Trace observes that most Tweets are of the topic *great* and *meet* in the land use *recreation* in New York City . In comparison they observe (for Atlanta), most Tweets are of topic *job* in the land use *work.* Exploring further, Trace considers a few topics such as *just*, *tn*, to be less comprehensive. They see an overview of the topic model showing other topics and their associated keywords in the Model Overview Panel (see Figure 1-B, C). In the interface they discard the shown topics and trigger InMacs to include other interesting topics by clicking the refresh button. In response, InMacs updates the heatmaps with a set of new topics.

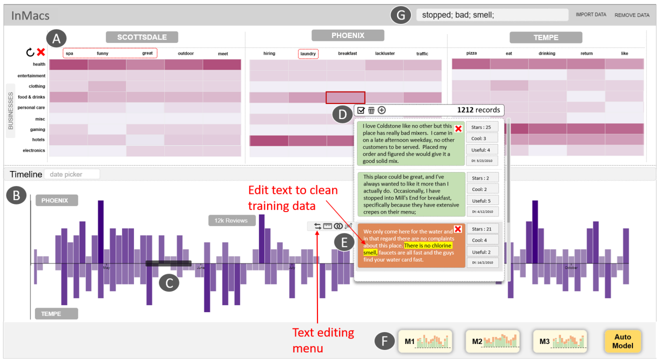
Next, Trace looks at the Timeline View to find the time and month of the year with the most Tweet activity. They notice while Atlanta shows an uniform distribution of Tweet activity through out the year, in New York City most of the Tweets are between *Apr* and *Nov.* Trace understands that this can be attributed to the respective weather conditions of the cities (New York City experiences extreme cold winters). Trace inspects the topics, *job*, and *work* to compare their land uses and the discussed keywords across the two cities. They find in New York City Tweets with *work* related topic are mostly discussed in the land use *recreation*, while in Atlanta *job* related Tweets are found in *work* land use (see Figure 1-A). Content with the analysis and insights gained so far, Trace proceeds to construct sentiment classifiers using the Auto Model button. They inspect the confusion matrix (see Figure 1-E) that shows the sentiments *Positive*, *Negative*, and *Neutral* and the prediction accuracy of 62% on test data. They also see that the heatmaps are now color encoded by the average sentiment score of Tweets per cell.

Trace then hovers overs the cell representing the topic *beer* in the land use *service* (see Figure 1-I). They inspect the word-cloud visualization (see Figure 1-C) to find keywords discussed in this topic and then the Timeline View to infer that most *beer* related Tweets were observed between the month *July-Sep.* Based on the red cross icon, they observe that the ground truth sentiment label for many Tweets are incorrect; which they immediately correct from the Text View (see Figure 1-H). Next Trace spot checks few other Tweets for both cities and triggers AutoML to construct a new sentiment classifier. They open the Model Overview Panel to see that the prediction accuracy on the test set increased to 83.34%. Trace also views the histogram view (see Figure 1-D) showing number of correctly predicted Tweets (by time) to find time ranges where incorrect predictions were mostly recorded. They click on these histogram bins to find that these Tweets contain the topic *job* with many keywords that are advertisement related. To filter such Tweets, Trace searches the keyword "Job Posting" on the search bar (see Figure 1-G). They discard these Tweets from the training set. Finally they trigger AutoML to build a new topic model and a new sentiment classifier (see Figure 1-F). They are pleased to see the prediction accuracy on test set jumped to 92.3%. Furthermore, from the Model Overview thumbnails they see the histogram charts (see Figure 1-F) to inspect the incremental progress they have made in improving the sentiment classifiers' accuracy over time. Content with the model traiing, Trace exports the data and the models for further analysis on the distinction between the urban characteristics of the two cities.

7 CASE STUDY 2: YELP REVIEWS

Yelp business review data [4] contains two hundred thirty thousand user reviews of various businesses (e.g., salons, cafes, hotels, etc.) on Yelp in the state of Arizona, USA. Using this data Shana, a business strategist seeks to compare peoples' sentiments on various businesses from Phoenix, Scottsdale, and Tempe, in Arizona, USA. Specifically they seek to compare: (1) peoples' sentiment and feedback in relation to various business categories, and (2) topics of discussion with time across various locations. The data contains thirty two attributes including *review-text*, *username*, *longitude*, *review-date*, *business-type*, *review-label(1-5)* (target variable), *business-city*, *business-categories*, *review-usefulness*, etc. Sentiments are annotated *positive* if the *review-label* is above 3, *negative* if below 3, and *neutral* if it equals 3. Shana splits the data into one fifty thousand training samples and loads it in InMacs to continue analysis.

Shana sees heatmap matrices for each of the three cities (see Figure 4-A), where the color on a cell reflects number of business review of a specific category and topic. For each matrix the rows are derived from the '*business-categories*' attribute in the data, e.g., *Health/Life-Activity*, *Arts Entertainment*, *Clothing*, *Food*, *Hotels*, *Misc*, etc. while the column represents topics derived from the text corpus. They hover the mouse on the category *Food* for the three cities to find various topics associated with food. They observe that the city “Tempe” shows a large number of similar topics related to “Mediterranean Food”. Further they look at the Timeline View (Figure 4-B) to find most of these reviews are posted between the month of *Mar-Jun* (Figure 4-C). Comparing the three heatmaps side by side, Shana notices that Phoenix has many negative reviews for the business category *Hotels*. Most of these reviews are of the topics “Laundry”, and “Breakfast”. To explore further, they open the Model Overview Panel to see other topics for these reviews. They select the topics “Cleanliness” and “Price” to be included in the heatmap matrix. Next they construct a new sentiment classifier by triggering the AutoML button (see Figure 4-F). They review the updated heatmap matrices which are now color encoded by the average sentiment score per cell of the user reviews. However, they find the sentiment classifiers’ performance is relatively poor on the test set (68.32%).



I love Coldstone like no other but this place has really bad mixers. I came in on a late afternoon weekday, no other customers to be served. Placed my order and figured she would give it a good solid mix.

This place could be great, and I've always wanted to like it more than I actually do. Occasionally, I have stopped into Mill's End for breakfast, specifically because they have extensive crepes on their menu;

We only come here for the water and in that regard there are no complaints about this place. There is no chlorine smell, faucets are all fast and the guys find your water card fast.

Figure 4: A. Heatmap comparison of three cities. B. Timeline view. C. Interactive brushing to filter data. D. Text view shown on mouse over. E. Edit text to add/remove/delete. F. Model detail view with AutoML button. G. Search bar.

Motivated to improve the performance of the classifier, Shana decides to spot-check few business reviews. From the Model Overview Panel they click on the cell of the confusion matrix that has the most number of incorrect predictions. All the views update to visually represent these data samples. From the topic-keyword scatterplot and the keyword bar chart view (Figure 1-B) they find that most of the poor predictions were because of the keywords “stopped”, “bad”, and “smell”. Next using the search bar (Figure 4-G), they specifically search for reviews with these keywords. From the Text View (Figure 4-D) they remove the negative keywords from these and other similar reviews that contains these words. They also notice that most of the positive reviews used these words at the end of their review. They specify to only use the first 100 words of such reviews (Figure 4-E). Next Shana constructs a new sentiment classifier.

InMacs responds back with a new sentiment classifier with an improved prediction accuracy of 83.4% on test set. Shana continues to spot check business reviews and predict sentiment labels, to find that most of the predictions are satisfactory. They compare the heatmaps across three cities to find that business related to Healthcare/Lifestyle category are booming in Scottsdale (based on the sudden increase of positively rated user reviews). Further, based on the “nightlife” keyword search they conclude that Phoenix is the best location for “nightlife” related business services, many of which received positive ratings. In addition, they find that out of the three cities, Tempe shows sudden spike in “Food” related reviews indicating the occurrence of festive events (e.g., feast furr carnival) in Tempe. However, there are many negative reviews for the business category *Hotel* in both Phoenix and Scottsdale, which may be a matter of concern. Finally, Shanaexports both the data and the model to continue their analysis.

8 DISCUSSION AND LIMITATION

Model hyperparameterizations: Through our design study with urban planners, we noticed often the problem and the questions they seek to solve, are ill-defined and not known apriori. This makes the model construction process partly exploratory in nature that leads to constructing models with varied model hyperparameterizations. However, we also observed that they were overwhelmed with the numerous variants of model hyperparameterizations. This motivated our design decision to minimize exposure of model hyperparameterizations and maximize interactive pre-processing of input data to construct models that better characterizes user goals. That being said, many visual elements in the tool encodes information pertaining to a ML models’ performance metrics such as *confusion matrix*, *accuracy/precision scores*, *classifier’s prediction probabilities* etc. which we realized from the conversations with the users that they are comfortable working with.

Consistency between human and machine: In a human in the loop based approach, ensuring consistency between users expectation and the ML model output is crucial but often gets lost. While the ML model is driven by underlying mathematical functions that seek to map the training data accurately, there may be external knowledge to the problem domain that is only known to the human. This discrepancy may cause an undesirable gap in what the user expects and what the model delivers. When such inconsistency persists, the human may not trust the model or may accept that their expected patterns do not exist in the data. In InMacs, we sought to minimize this by allowing urban planners to externalize their knowledge by directly updating the training data, that drives the ML models’ logical reasoning. We strived to provide users complete agency and control to adjust, augment, and pre-process data as an active participant in the modeling pipeline as opposed to merely be a data resource for any AutoML solver.

9 CONCLUSION

In this paper, we empower domain experts with limited (or no) expertise in ML to actively construct ML models in the process of data exploration. Specifically, we work closely with urban planners to help them to: (1) compare peoples’ sentiment and topic of discussion across various geolocations, (2) interactively construct sentiment classifiers with large scale social media data, and (3) assert/frame critical domain-specific hypotheses using AutoML. We present a novel VA system, InMacs that combines interactive visualizations with an AutoML model solver to help urban planners construct sentiment classifiers and topic models on sequence data from Twitter API. Furthermore, through another use case, we show the generalizability of the system and our technique on any text-based input (e.g, Yelp business review data). We allow users to actively adjust models using a visual interface with an AutoML solver.

REFERENCES

[1] 2014. *Statistical Modeling by Gesture: A graphical, browser-based statistical interface for data repositories.* http://ceur-ws.org/Vol-1210/datawiz2014_05.pdf

[2] 2020. 200,000 USA geolocated tweets. Free Twitter Dataset. http://followthehashtag.com/datasets/free-twitter-dataset-usa-200000-free-usa-tweets/. Accessed: 2020-05-18.

[3] 2020. Tentative Property Assessment Data. https://www1.nyc.gov/site/finance/taxes/property-assessments.page. Accessed: 2020-05-18.

[4] 2020. Yelp Dataset. https://data.world/brianray/yelp-reviews. Accessed: 2020-05-18.

[5] S. Hassan Ameli, Shima Hamidi, Andrea Garfinkel-Castro, and Reid Ewing. 2015. Do Better Urban Design Qualities Lead to More Walking in Salt Lake City, Utah? *Journal of Urban Design* 20, 3 (2015), 393–410. https://doi.org/10.1080/13574809.2015.1041894 arXiv:https://doi.org/10.1080/13574809.2015.1041894

[6] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine* 35, 4 (2014), 105–120.

[7] Saleema Amershi, James Fogarty, Ashish Kapoor, and Desney S Tan. 2011. Effective End-User Interaction with Machine Learning.. In *AAAI*.

[8] Saleema Amershi, James Fogarty, and Daniel Weld. 2012. Regroup: Interactive Machine Learning for On-demand Group Creation in Social Networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 21–30. https://doi.org/10.1145/2207676.2207680

[9] B. Bach, E. Pietriga, and J. Fekete. 2014. GraphDiaries: Animated Transitions andTemporal Navigation for Dynamic Networks. *IEEE Transactions on Visualization and Computer Graphics* 20, 5 (2014), 740–754.

[10] James Bergstra, Dan Yamins, and David D Cox. 2013. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. In *Proceedings of the 12th Python in Science Conference*. 13–20.

[11] J. Bergstra, D. Yamins, and D. D. Cox. 2013. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28 (ICML'13)*. JMLR.org, I–115–I–123.

[12] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3, null (March 2003), 993–1022.

[13] Shelley Boulianne. 2015. Social media use and participation: a meta-analysis of current research. *Information, Communication & Society* 18, 5 (2015), 524–538. https://doi.org/10.1080/1369118X.2015.1008542 arXiv:https://doi.org/10.1080/1369118X.2015.1008542

[14] S. Bremm, T. von Landesberger, M. Heß, T. Schreck, P. Weil, and K. Hamacherk. 2011. Interactive visual comparison of multiple trees. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 31–40.

[15] Dylan Cashman, Shah Rukh Humayoun, Florian Heimerl, Kendall Park, Subhajit Das, John Thompson, Bahador Saket, Abigail Mosca, John T. Stasko, Alex Endert, Michael Gleicher, and Remco Chang. 2019. A User-based Visual Analytics Workflow for Exploratory Model Analysis. *Comput. Graph. Forum* 38, 3 (2019), 185–199. https://doi.org/10.1111/cgf.13681

[16] Marco Cavallo and Çagatay Demiralp. 2018. Clustrophile 2: Guided Visual Clustering Analysis. *IEEE Transactions on Visualization and Computer Graphics* 25 (2018), 267–276.

[17] Das, Cashman Subhajit, Chang Dylan, Remco, Endert, and Alex. 2019. BEAMES: Interactive Multi-Model Steering, Selection, and Inspection for Regression Tasks. In *IEEE CGA*.

[18] Nicholas Diakopoulos, Dag Elgesem, Andrew Salway, Amy Zhang, and Knut Hofl. 2015. Compare clouds: Visualizing text corpora to compare media frames. In *In Proc. of IUI Workshop on Visual Text Analytics*.

[19] Dennis Dingen, Marcel van 't Veer, Patrick Houthuizen, Eveline H. J. Mestrom, Hendrikus H. M. Korsten, Arthur R. A. Bouwman, and Jarke J. van Wijk. 2018. RegressionExplorer: Interactive Exploration of Logistic Regression Models with Subgroup Analysis. *IEEE Transactions on Visualization and Computer Graphics* 25 (2018), 246–255.

[20] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. 2019. Neural Architecture Search: A Survey. *J. Mach. Learn. Res.* 20 (2019), 55:1–55:21. http://jmlr.org/papers/v20/18-598.html

[21] A. Endert, W. Ribarsky, C. Turkay, B.L. William Wong, I. Nabney, I. Díaz Blanco, and F. Rossi. 2017. The State of the Art in Integrating Machine Learning into Visual Analytics. *Computer Graphics Forum* (2017). https://doi.org/10.1111/cgf.13092

[22] Jerry Alan Fails and Dan R. Olsen. 2003. Interactive Machine Learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces (IUI '03)*. Association for Computing Machinery, New York, NY, USA, 39–45. https://doi.org/10.1145/604045.604056

[23] Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, and Frank Hutter. 2015. Efficient and robust automated machine learning. In *Advances in Neural Information Processing Systems*. 2962–2970.

[24] Yolanda Gil, James Honaker, Shikhar Gupta, Yibo Ma, Vito D'Orazio, Daniel Garijo, Shruti Gadewar, Qifan Yang, and Neda Jahanshad. 2019. Towards Human-Guided Machine Learning. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. Association for Computing Machinery, New York, NY, USA, 614–624. https://doi.org/10.1145/3301275.3302324

[25] Michael Gleicher, Danielle Albers, Rick Walker, Ilir Jusufi, Charles D. Hansen, and Jonathan C. Roberts. 2011. Visual Comparison for Information Visualization. *Information Visualization* 10, 4 (Oct. 2011), 289–309. https://doi.org/10.1177/1473871611416549

[26] Samiul Hasan and Satish V. Ukkusuri. 2014. Urban activity pattern classification using topic models from online geo-location data. *Transportation Research Part C: Emerging Technologies* 44 (2014), 363 – 381. https://doi.org/10.1016/j.trc.2014.04.003

[27] Danny Holten and Jarke J. van Wijk. 2008. Visual Comparison of Hierarchically Organized Data. In *Proceedings of the 10th Joint Eurographics / IEEE - VGTC Conference on Visualization (EuroVis'08)*. The Eurographs Association John Wiley Sons, Ltd., Chichester, GBR, 759–766. https://doi.org/10.1111/j.1467-8659.2008.01205.x

[28] Andreas Holzinger. 2016. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics* 3, 2 (01 Jun 2016), 119–131. https://doi.org/10.1007/s40708-016-0042-6

[29] Zhenjun Hu, Tianhua Frith, Niu, and Zhiping Weng. 2003. SeqVISTA: a graphical tool for sequence feature visualization and comparison. *BMC Bioinformatics* 4 (2003), 1471–2105. https://doi.org/10.1145/2702123.2702419

[30] Stefan Jänicke, Judith Blumenstein, Michaela Rücker, Dirk Zeckzer, and Gerik Scheuermann. 2018. TagPies: Comparative Visualization of Textual Data. In *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2018) - Volume 3: IVAPP, Funchal, Madeira, Portugal, January 27-29, 2018*, Alexandru Telea, Andreas Kerren, and José Braz (Eds.). SciTePress, 40–51. https://doi.org/10.5220/0006548000400051

[31] Markus John, Eduard Marbach, Steffen Lohmann, Florian Heimerl, and Thomas Ertl. 2018. MultiCloud: Interactive Word Cloud Visualization for the Analysis of Multiple Texts. In *Proceedings of the 44th Graphics Interface Conference (GI '18)*. Canadian Human-Computer Communications Society, Waterloo, CAN, 34–41. https://doi.org/10.20380/GI2018.06

[32] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1746–1751. https://doi.org/10.3115/v1/D14-1181

[33] Felix Kling and Alexei Pozdnoukhov. 2012. When a City Tells a Story: Urban Topic Analysis. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '12)*. Association for Computing Machinery, New York, NY, USA, 482–485. https://doi.org/10.1145/2424321.2424395

[34] Brent Komer, James Bergstra, and Chris Eliasmith. 2014. Hyperopt-sklearn: automatic hyperparameter configuration for scikit-learn. In *ICML workshop on AutoML*.

[35] Lars Kotthoff, Chris Thornton, Holger H Hoos, Frank Hutter, and Kevin Leyton-Brown. 2016. Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. *Journal of Machine Learning Research* 17 (2016), 1–5.

[36] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura E. Barnes, and Donald E. Brown. 2019. Text Classification Algorithms: A Survey. (2019). http://arxiv.org/abs/1904.08067 cite arxiv:1904.08067.

[37] Da Kuang, Jaegul Choo, and Haesun Park. 2014. Nonnegative Matrix Factorization for Interactive Topic Modeling and Document Clustering.

[38] Po-Ming Law, Subhajit Das, and Rahul C. Basole. 2019. Comparing Apples and Oranges: Taxonomy and Design of Pairwise Comparisons within Tabular Data. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300409

[39] Dawei Li, Yujia Zhang, and Cheng Li. 2019. Mining Public Opinion on Transportation Systems Based on Social Media Data. *Sustainability* 11, 15 (2019). https://doi.org/10.3390/su11154016

[40] F.-F. Li and J. Li. [n.d.]. Cloud automl: Making ai accessible to every business. https://www.blog.google/topics/google-cloud/cloud-automl-making-ai-accessible-every-business/. Accessed: 2019-06-30.

[41] Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Morgan Claypool Publishers.

[42] X. Liu, A. Xu, L. Gou, H. Liu, R. Akkiraju, and H. Shen. 2016. SocialBrands: Visual analysis of public perceptions of brands on social media. In *2016 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 71–80.

[43] Érick Lopez-Ornelas, Rocío Abascal-Mena, and Sergio Zepeda-Hernández. 2017. SOCIAL MEDIA PARTICIPATION IN URBAN PLANNING: A NEW WAY TO INTERACT AND TAKE DECISIONS.

[44] Michael E. Martin and Nadine Schuurman. 2017. Area-Based Topic Modeling and Visualization of Social Media for Qualitative GIS. *Annals of the American Association of Geographers* 107 (2017), 1028 – 1039.

[45] Tamara Munzner, François Guimbretière, Serdar Tasiran, Li Zhang, and Yunhong Zhou. 2003. TreeJuxtaposer: Scalable Tree Comparison Using Focus+Context with Guaranteed Visibility. In *ACM SIGGRAPH 2003 Papers (SIGGRAPH '03)*. Association for Computing Machinery, New York, NY, USA, 453–462. https://doi.org/10.1145/1201775.882291

[46] Sander Münster, Christopher Georgi, Katrina Heijne, Kevin Klamert, Jörg [Rainer Noennig], Matthias Pump, Benjamin Stelzle, and Han [van der Meer]. 2017. How to involve inhabitants in urban design planning by using digital tools? An overview on a state of the art, key challenges and promising approaches. *Procedia Computer Science* 112 (2017), 2391 – 2405. https://doi.org/10.1016/j.procs.2017.08.102 Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 21st International Conference, KES-20176-8 September 2017, Marseille, France.

[47] Er Pak and Patrick Paroubek. [n.d.]. Twitter as a corpus for sentiment analysis and opinion mining. In *In Proceedings of the Seventh Conference on International Language Resources and Evaluation*.

[48] Stelios Paparizos, Jignesh M Patel, and HV Jagadish. 2007. SIGOPT: Using schema to optimize XML query processing. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. IEEE, 1456–1460.

[49] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[50] Debjyoti Paul, Feifei Li, Murali Krishna Teja, Xin Yu, and Richie Frost. 2017. Compass: Spatio Temporal Sentiment Analysis of US Election What Twitter Says!. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. ACM, New York, NY, USA, 1585–1594. https://doi.org/10.1145/3097983.3098053

[51] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[52] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162

[53] Richard A. Plunz, Yijia Zhou, Maria Isabel [Carrasco Vintimilla], Kathleen Mckeown, Tao Yu, Laura Uguccioni, and Maria Paola Sutto. 2019. Twitter sentiment in New York City parks as measure of well-being. *Landscape and Urban Planning* 189 (2019), 235 – 246. https://doi.org/10.1016/j.landurbplan.2019.04.024

[54] Foster Provost and Ron Kohavi. 1998. On Applied Research in Machine Learning. In *Machine learning*. 127–132.

[55] Helen Roberts, Jon Sadler, and Lee Chapman. 2019. The value of Twitter data for determining the emotional responses of people to urban green spaces: A case study and critical evaluation. *Urban Studies* 56, 4 (2019), 818–835. https://doi.org/10.1177/0042098017748544 arXiv:https://doi.org/10.1177/0042098017748544

[56] Dominik Sacha, Michael Sedlmair, Leishi Zhang, John Aldo Lee, Jaakko Peltonen, Daniel Weiskopf, Stephen C. North, and Daniel A. Keim. 2017. What you see is what you can change: Human-centered machine learning by interactive visualization. *Neurocomputing* 268 (2017), 164–175.

[57] Aecio Santos, Sonia Castelo, Cristian Felix, Jorge Piazentin Ono, Bowen Yu, Sungsoo Hong, Claudio T. Silva, Enrico Bertini, and Juliana Freire. 2019. Visus: An interactive system for automatic machine learning model building and curation.

[58] Chris Thornton, Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. 2013. Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 847–855.

[59] Stef Van Den Elzen and Jarke J van Wijk. 2011. Baobabview: Interactive construction and analysis of decision trees. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*. IEEE, 151–160.

[60] Yong Wang, Hammad Haleem, Conglei Shi, Yanhong Wu, Xun Zhao, Siwei Fu, and Huamin Qu. 2018. Towards Easy Comparison of Local Businesses Using Online Reviews. *Computer Graphics Forum* 37, 3 (2018), 63–74. https://doi.org/10.1111/cgf.13401 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13401

[61] Y. Wu, Z. Chen, G. Sun, X. Xie, N. Cao, S. Liu, and W. Cui. 2018. StreamExplorer: A Multi-Stage System for Visually Exploring Events in Social Streams. *IEEE Transactions on Visualization and Computer Graphics* 24, 10 (2018), 2758–2772.

[62] J. Xu, Y. Tao, H. Lin, Rongjie Zhu, and Yuyu Yan. 2017. Exploring controversy via sentiment divergences of aspects in reviews. In *2017 IEEE Pacific Visualization Symposium (PacificVis)*. 240–249.

[63] Shanqi Zhang and Rob Feick. 2016. Understanding Public Opinions from Geosocial Media. *ISPRS International Journal of Geo-Information* 5, 6 (2016).

[64] Jian Zhao, Zhicheng Liu, Mira Dontcheva, Aaron Hertzmann, and Alan Wilson. 2015. MatrixWave: Visual Comparison of Event Sequence Data. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 259–268. https://doi.org/10.1145/2702123.2702419