Supplementary Material

Appendix 1: iterative filtering process to identify football-related tweets

The graphical illustration of the filtering process is shown in Figure 1.

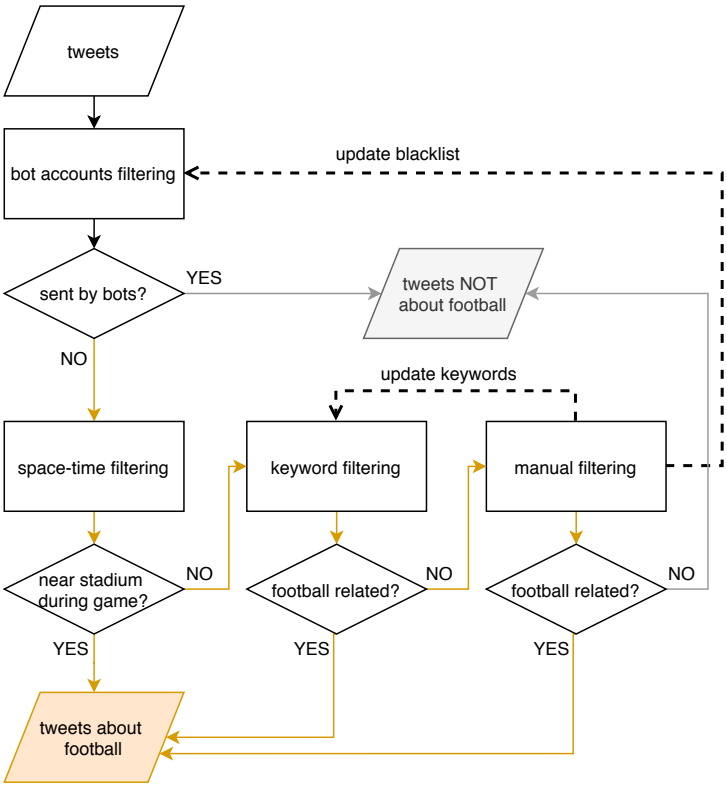

Figure 1 Iterative filtering process for identifying football-related tweets.

The iterative workflow starts with a subtractive filtering process that removes spam, bot, and cyborg tweets (Tsou *et al.* 2017). The remaining tweets are subjected to an additive filtering process to identify football-related tweets using three separate filters: a space-time window filter, a keyword filter, and a manual inspection. The three filters are implemented in sequence, and tweets that satisfy any of the three filters will be added into the final pool of football-related tweets.

The space-time window filter identifies tweets that were geotagged at Beaver Stadium within a temporal range of three hours before and after the game event. The keyword filter identifies football-related tweets by recognizing keywords associated with football events. The remaining unidentified tweets are then subjected to a manual inspection conducted by an

analyst to recognize football-related tweets. During the manual inspection process, the analyst also updates the blacklist of bot accounts used in the subtractive filtering process and updates the keyword filter.

This iterative workflow was applied to Twitter data in batches, to increase the accuracy and efficiency of identifying football-related tweets with each iteration. The specific criteria of each filter and the number of tweets filtered are shown in Table 1.

Table 1 Statistics of filtered tweets

Type of filtering	**Examples**	**Percentage of tweets identified by the filter**	**Number of tweets filtered**
bot accounts filtering	1. Tweeting platform is one of the following: 'TweetMyJOBS', 'circlepix', 'Tweetbot for Mac'. 2. Username includes one of the following words 'career', 'job', 'transport', or 'beer'. Tweet text includes hashtags 'job' or 'hiring'. 3. Organizational accounts such as 'SOTAwatch spots', 'TTN Harrisburg', 'Spotter Network Inc'.	20.3%	1032
space-time filtering	tweets that are inside the Beaver Stadium and posted less than three-hour difference of the game time	7.2%	366
keyword filtering	Examples of keywords are 'psufootball', 'pennstatefootball', 'pennstatefball', '#WeAre', 'game day', 'Beaver Stadium', 'vsPSU', 'home game', 'homecoming', 'go state', 'game time', 'tailgat', '#GoBigBlue', '#psuunrivaled', 'B1G', 'Penn State Nittany Lions vs.' 'go lions', '#rosebowl', 'touchdown', 'quarterback', 'stripeout', 'whiteout'.	26.7%	1354
manual filtering	based on external link to pictures, emoji, and meaning of the tweet	2.4%	120

The iterative filtering process classified all tweets into three categories: (1) football-related tweets, (2) other tweets sent by humans but not related to football, and (3) noise sent by spam, bot, and cyborg accounts. We calculated the volume of tweets by game weekends and non-game weekends for these three categories to understand the variations in the tweeting patterns (Figure 2).

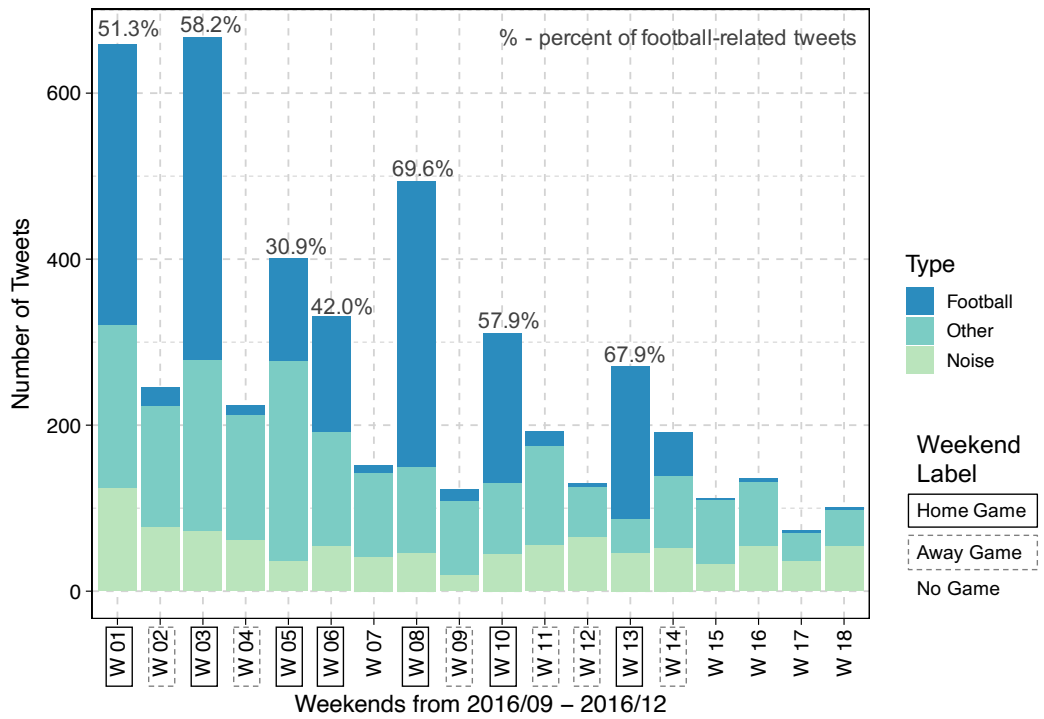

Figure 2 Number of tweets by category (football-related tweets, other tweets sent by people, and noise sent by bot accounts) for 18 weekends from Sept. to Dec. 2016. The labels of home game weekends, away game weekends, and no-game weekends are marked respectively with a solid border, a dashed border, and no border.

Appendix 2: inferring home locations of Twitter users using geolocated timeline tweets

Identifying a user's home location is a four-step process: (1) Geolocated timeline tweets for each football fan are extracted and chronologically ordered. (2) Consecutive tweets from the same place are grouped together and considered as one visit to that place. (3) Total time spent during each visit is calculated by counting the number of days between the first and last geolocated tweets (This calculation may exaggerate duration at frequented locations if there is intervening travel with no Twitter posts on any day during that travel). (4) The time spent during all visits to a place is summed up to calculate the accumulated duration of stay. The top three places with longest accumulated duration are selected as candidates for home locations.

To improve the accuracy of identifying home locations, the top three candidates were

compared with the user-identified profile location to jointly decide the home location. Our analysis shows that 48% of users in our data set have a valid profile location as precise as a county or city. The criteria used to assign home locations are listed below:

- If the user's profile location matches the place with the most accumulated time, then that location is considered to be the home location.
- If the accumulated time at the place with the most time is longer than one month (Yin *et al.* 2016) and more than double the time spent at the place with the second most time, then the home location is the place with the most accumulated time.
- If neither of the above is met, then we manually check the tweeting history for the remaining users to identify the most likely home location. When this step is needed, multiple possible clues are used, including places named in tweets, the content of images in tweets, etc.

Appendix 3: classifying topics of tweets using Labeled LDA model

The Labeled LDA model is a generative probabilistic model that 'describes a (theoretical) process of generating labeled documents' (Ramage *et al.* 2009). The process consists of two main steps: (1) for each topic k, the process generates a multinomial distribution β_k over all words from a Dirichlet prior η, (2) for each document d, the process then generates a multinomial distribution $\theta^{(d)}$ over all topics from a Dirichlet prior α. The topic-document assignment $\theta^{(d)}$ is restricted to the labels of the document. The implementation of the Labeled LDA method is to use a collection of labeled documents to train a model that best describes the word and label distributions observed in the data set. The process is conducted by running a Gibbs sampling method (Griffiths and Steyvers 2004).

To produce a labeled training data set as input for the Labeled LDA model, we used a free text tagging service, Open Calais API. Open Calais API has a rate limit of 5,000 requests

per day, which makes it time-consuming to label all users' timeline tweets[1]. Thus, a subset of 60,000 timeline tweets, approximately 38% of all geolocated timeline tweets, were submitted to Open Calais API for labeling. The text was cleaned by removing emoji and URL links before the labeling process.

Once the Labeled LDA model learned the word distribution for each label through the training process, we predicted new labels for untagged tweets based on the word-per-label distribution. The top 20 words and their associated probabilities for each label are shown in Table 3. We used 10-fold cross-validation to evaluate the model. In each round of cross-validation, 90% of the labeled tweets were used as training data, and the remaining 10% of the labeled tweets were used as test data. During the testing phase, the model predicts a list of labels for each test tweet based on the word-per-label distribution. The predicted labels are ranked from most to least likely. A comparison is conducted between the predicted labels and the true labels tagged by the Open Calais API to evaluate the performance of the prediction.

Four metrics were used to measure the performance (Powers 2011, Schröder 2018): AUC-ROC curve (Area Under the Curve-Receiver Operating Characteristics), one error, two error, and F1 score (macro average). The explanations for the four measures are listed below:

- AUC-ROC curve (Area Under the Curve-Receiver Operating Characteristics): The metric measures the ability of the model to distinguish different classes. AUC-ROC ranges from 0 to 1, and higher value indicates better separability among classes.
- One error: This metric is a measure of the accuracy of the prediction. It calculates the percentage of correctly labeled tweets by only considering the highest probable label predicted by Labeled LDA.

1 The Open Calais service daily capacity has decreased to a maximum of 500 documents per day starting from March 31st, 2020.

- Two error: This metric is similar to the one error metric, but considers the top two highest probable labels instead of one.
- F1 score (macro average): The metric measures a test's accuracy. It is a harmonic weighted average of the precision and recall of a classification. The value ranges from 0 to 1, and higher value indicates a better performance of the classification.

The final measured performance was averaged across the 10-fold cross-validation. Results show that our model successfully separates different classes; thus, it distinguishes among the prototypical topics (Table 2). Yet, many individual entities are hard to classify. The approach can often narrow things down to a tweet belonging to one of two categories, but does not do as well in making the final distinction between them. In addition, we conducted a validation process by asking two analysts to label 2000 randomly sampled tweets manually. The macro F1 score from the two analysts is 0.812, which is slightly lower than the F1 score 0.867 of the Labeled LDA method. The result demonstrates the inherent difficulty in labeling tweets and that the accuracy of the classification model is higher than manual tagging.

Table 2 Measured performance of the Labeled LDA model.

Metrics	**Measurement**
AUC ROC	0.948
one error	0.748
two error	0.854
F1 score (macro average)	0.867

Table 3 Top 20 words predicted by Labeled LDA model for each label

-- label 1: Social Issues	-- label 2: War_Conflict	-- label 3: Religion_Belief	-- label 4: Human Interest	-- label 5: Health_Medical_Pharma	-- label 6: Other
hate: 0.0628 people: 0.0282 penn: 0.0209 life: 0.0193 im: 0.0193 social: 0.0185 tulum: 0.0161 hope: 0.0153 wait: 0.0137 do: 0.0137 medium: 0.0129 suicide: 0.0129 mexico: 0.0113 texas: 0.0089 house: 0.0089 dying: 0.0073 live: 0.0073 aint: 0.0065 obama: 0.0065 gay: 0.0056	airport: 0.0579 international: 0.0307 charleston: 0.0273 dfw: 0.0205 lga: 0.0171 laguardia: 0.0171 ch: 0.0137 sc: 0.0137 tx: 0.0102 im: 0.0102 worth: 0.0102 dallasfort: 0.0102 yorktown: 0.0068 aircraft: 0.0034 carrier: 0.0034 hartsfieldjackson: 0.0034 mco: 0.0034 chsdca: 0.0034 orlando: 0.0034 miracle: 0.0034	god: 0.0763 life: 0.0456 word: 0.0163 literally: 0.0156 holy: 0.0144 jesus: 0.0131 truth: 0.0119 im: 0.0100 people: 0.0094 saint: 0.0088 bless: 0.0088 temple: 0.0081 church: 0.0081 lord: 0.0075 prayer: 0.0069 louis: 0.0069 hour: 0.0063 bagel: 0.0056 understand: 0.0050 true: 0.0050	love: 0.0287 happy: 0.0232 birthday: 0.0211 mom: 0.0197 amaze: 0.0166 friend: 0.0155 hope: 0.0154 park: 0.0130 dog: 0.0126 im: 0.0109 ill: 0.0109 day: 0.0097 photo: 0.0096 amp: 0.0088 baby: 0.0082 drink: 0.0081 youre: 0.0080 do: 0.0070 laugh: 0.0069 life: 0.0069	wait: 0.0372 im: 0.0323 drink: 0.0316 blue: 0.0150 care: 0.0133 heart: 0.0126 park: 0.0110 brain: 0.0093 blood: 0.0080 colorado: 0.0077 fat: 0.0077 walk: 0.0073 lunch: 0.0073 dr: 0.0073 la: 0.0070 tea: 0.0067 pa: 0.0063 football: 0.0057 accurate: 0.0057 university: 0.0053	amp: 0.0615 im: 0.0383 fl: 0.0260 savannah: 0.0219 drink: 0.0219 dad: 0.0205 cream: 0.0178 morris: 0.0164 grill: 0.0164 tap: 0.0164 orlando: 0.0150 ice: 0.0109 ga: 0.0109 vista: 0.0096 lake: 0.0096 buena: 0.0096 park: 0.0082 cookie: 0.0082 magic: 0.0082 chapman: 0.0055
-- label 7: Law_Crime	**- label 8: Sports**	**-- label 9: Environment**	**-- label 10: Business_Finance**	**-- label 11: Disaster_Accident**	**-- label 12: Weather**
girl: 0.0576 im: 0.0488 san: 0.0440 ca: 0.0424 francisco: 0.0248 boy: 0.0184 airport: 0.0176 international: 0.0136 diego: 0.0136 car: 0.0112 hate: 0.0112 phoenix: 0.0088 beer: 0.0064 juan: 0.0064 drink: 0.0056 penalty: 0.0048 city: 0.0048 california: 0.0048 harbor: 0.0040 law: 0.0040	game: 0.0556 team: 0.0245 win: 0.0232 play: 0.0210 stadium: 0.0187 football: 0.0161 im: 0.0130 chicago: 0.0096 field: 0.0087 coach: 0.0081 fan: 0.0067 favorite: 0.0067 final: 0.0066 ball: 0.0064 beaver: 0.0063 watch: 0.0062 goal: 0.0059 penn: 0.0057 airport: 0.0056 psu: 0.0052	brew: 0.0483 drink: 0.0424 im: 0.0424 company: 0.0342 lake: 0.0318 fl: 0.0208 buena: 0.0204 vista: 0.0204 ocean: 0.0133 city: 0.0094 ice: 0.0082 island: 0.0063 ale: 0.0059 season: 0.0055 pic: 0.0055 keuka: 0.0055 tree: 0.0047 revelation: 0.0043 md: 0.0043 ipa: 0.0043	im: 0.0485 market: 0.0249 york: 0.0230 spend: 0.0151 ny: 0.0138 hotel: 0.0112 business: 0.0092 vega: 0.0085 las: 0.0085 airport: 0.0085 beer: 0.0085 international: 0.0079 company: 0.0072 bank: 0.0072 mall: 0.0066 car: 0.0066 nv: 0.0066 denver: 0.0059 read: 0.0059 customer: 0.0053	drink: 0.0931 im: 0.0389 ny: 0.0237 york: 0.0213 food: 0.0163 city: 0.0152 bad: 0.0139 fire: 0.0115 amp: 0.0091 weather: 0.0088 dead: 0.0088 airport: 0.0088 nurse: 0.0088 las: 0.0085 driver: 0.0085 ipa: 0.0081 valley: 0.0075 vega: 0.0071 international: 0.0064 storm: 0.0061	tomorrow: 0.0535 snow: 0.0535 weather: 0.0384 drink: 0.0361 photo: 0.0361 rain: 0.0337 weekend: 0.0326 degree: 0.0151 cold: 0.0105 warm: 0.0093 storm: 0.0082 winter: 0.0082 september: 0.0070 thursday: 0.0070 coast: 0.0058 shower: 0.0058 wind: 0.0047 east: 0.0047 la: 0.0047 fall: 0.0047

-- label 13: Technology_Internet	-- label 14: Labor	-- label 15: Politics	-- label 16: Education	-- label 17: Entertainment_Culture	-- label 18: Hospitality_Recreation
video: 0.0225 text: 0.0217 amp: 0.0208 moon: 0.0199 facebook: 0.0147 cafe: 0.0139 brewery: 0.0130 drink: 0.0121 apple: 0.0104 email: 0.0095 internet: 0.0087 window: 0.0078 website: 0.0078 google: 0.0078 iphone: 0.0078 post: 0.0078 game: 0.0069 phone: 0.0069 launch: 0.0061 system: 0.0052	job: 0.0378 peer: 0.0114 specialist: 0.0114 certify: 0.0114 weekendcoffeeshare: 0.0076 recovery: 0.0076 employee: 0.0076 price: 0.0038 workforce: 0.0038 generously: 0.0038 minimum: 0.0038 wage: 0.0038 retrain: 0.0038 initiative: 0.0038 kudo: 0.0038 development: 0.0038 bubbakoos: 0.0038 amen: 0.0038 everdelightful: 0.0038	party: 0.0471 im: 0.0228 berlin: 0.0182 york: 0.0159 ny: 0.0144 vote: 0.0137 house: 0.0129 airport: 0.0121 harlem: 0.0106 east: 0.0099 washington: 0.0091 office: 0.0084 jet: 0.0068 park: 0.0068 president: 0.0068 clinton: 0.0061 france: 0.0061 vacay: 0.0053 elmhurst: 0.0053 laguardia: 0.0053	school: 0.0851 college: 0.0756 happy: 0.0694 birthday: 0.0659 class: 0.0345 im: 0.0338 university: 0.0237 pa: 0.0195 kid: 0.0119 student: 0.0112 start: 0.0094 bro: 0.0087 parent: 0.0084 park: 0.0080 teacher: 0.0066 week: 0.0066 austin: 0.0066 homework: 0.0066 amp: 0.0063 homie: 0.0063	love: 0.0701 im: 0.0239 watch: 0.0210 disney: 0.0167 gonna: 0.0149 tonight: 0.0130 music: 0.0128 disneys: 0.0101 movie: 0.0098 song: 0.0088 guy: 0.0088 girl: 0.0085 world: 0.0084 live: 0.0073 rock: 0.0072 kingdom: 0.0067 birthday: 0.0067 miami: 0.0066 happy: 0.0064 friend: 0.0062	beach: 0.0357 drink: 0.0307 im: 0.0171 bar: 0.0138 amp: 0.0113 night: 0.0074 photo: 0.0072 day: 0.0066 london: 0.0062 park: 0.0060 hotel: 0.0058 dinner: 0.0053 resort: 0.0053 wine: 0.0051 ale: 0.0051 lake: 0.0050 beautiful: 0.0048 beer: 0.0048 restaurant: 0.0048 grill: 0.0047

Appendix 4

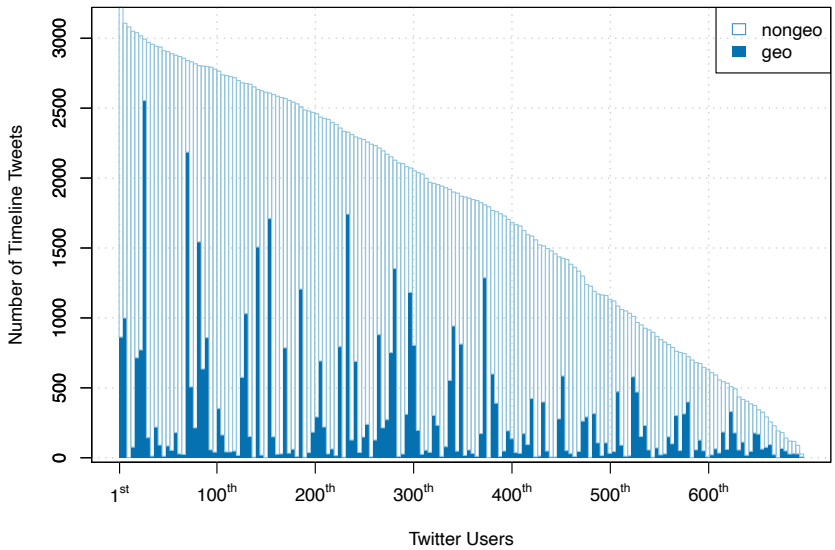

Figure 3 The number of timeline tweets by individual users in descending order (from 1st to 696th user). The timeline tweets were classified into geolocated tweets (dark blue) and non-geolocated tweets (light blue).

Appendix 5

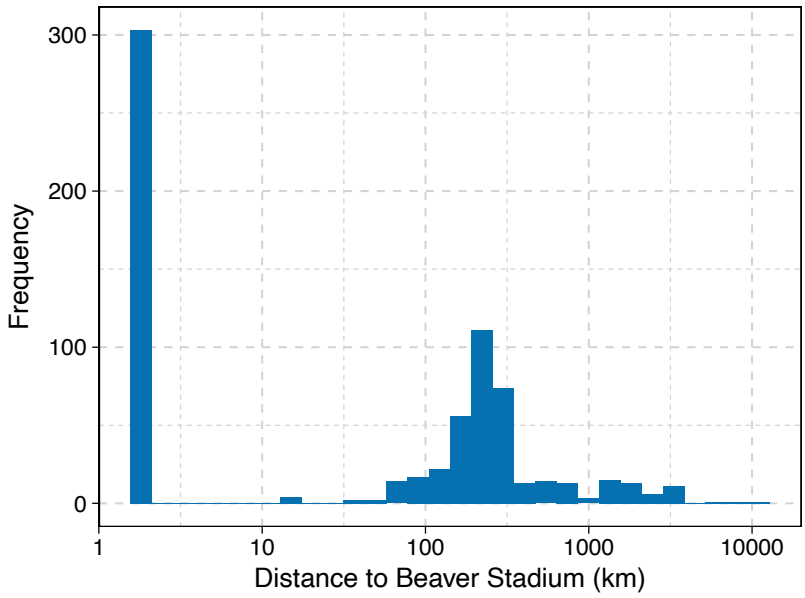

Figure 4 Histogram of the distance from football fan's home location to Beaver Stadium. The x-axis is on a logarithmic scale.

References

Griffiths, T.L. and Steyvers, M., 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101 (Supplement 1), 5228–5235.

Powers, D.M., 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

Ramage, D., Hall, D., Nallapati, R., and Manning, C.D., 2009. Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-labeled Corpora. *In*: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*. Stroudsburg, PA, USA: Association for Computational Linguistics, 248–256.

Schröder, K., 2018. Hierarchical Multiclass Topic Modelling with Prior Knowledge.

Tsou, M.-H., Zhang, H., and Jung, C.-T., 2017. Identifying Data Noises, User Biases, and System Errors in Geo-tagged Twitter Messages (Tweets). *arXiv:1712.02433 [cs]*.

Yin, J., Gao, Y., Du, Z., and Wang, S., 2016. Exploring Multi-Scale Spatiotemporal Twitter User Mobility Patterns with a Visual-Analytics Approach. *ISPRS International Journal of Geo-Information*, 5 (10), 187.