Constructing Accurate Confidence Intervals when Aggregating Social Media Data for Public Health Monitoring

Ashlynn R. Daughton[1,2*] and **Michael J. Paul**[1]
[1] Information Science, University of Colorado, Boulder, CO 80309
[2] Analytics, Intelligence, and Technology, Los Alamos National Laboratory, NM 87545
Ashlynn.Daughton@colorado.edu, Michael.J.Paul@colorado.edu

Abstract

Social media data are widely used to infer health related information (e.g., the number of individuals with symptoms). A typical approach is to use a machine learning classification to aggregate and count the information of interest. However, this approach fails to account for errors made by the classifier. This paper summarizes data mining concepts that account for classifier error when counting data instances, and then extends these ideas to propose a new algorithm for constructing confidence intervals of social media estimates that we show to be substantially more accurate than standard approaches on two influenza-related Twitter datasets.

Introduction

Social media posts have been used to infer trends related to a wide variety of health applications. A common approach to extract signals from social media is to first filter the data for relevant content, usually involving a combination of simple search queries and machine learning classification, and then aggregating the content by counting the number of relevant posts within specified groups (e.g., counts by week or by location) (Paul and Dredze 2017). This approach has been applied to influenza surveillance (Culotta 2010; Doan, Ohno-Machado, and Collier 2012), measuring vaccination attitudes (Mitra, Counts, and Pennebaker 2016) and behavior (Huang et al. 2017), and monitoring public health concerns (Ji, Chun, and Geller 2013).

A flaw in this approach is that the aggregated counts typically do not account for biases and errors introduced by the relevance filtering and classification step. While studies will typically report evaluation metrics of the accuracy of this step, once the accuracy is deemed "good enough", downstream statistical analysis is applied to the classified data and relevance classifications are treated as correct. Since almost all methods of filtering and classification will introduce some degree of error, we seek to better understand the effect this error has on downstream aggregation.

In the data mining community, the task of aggregating individually-classified instances is known as *quantification*, and various methods have been proposed to adjust for classification error to produce more accurate counts (Forman 2008). However, most social media studies do not draw on methods from the quantification literature when conducting statistical analyses of aggregated data, and to the best of our knowledge, these methods have not been applied to social media studies in the health domain.

The purpose of this short paper is to introduce concepts of quantification from the data mining community to the social media monitoring community; additionally, we present a new algorithm for constructing confidence intervals of social media estimates that we show to be more accurate than standard quantification approaches, as existing quantification techniques have been focused on point estimates rather than confidence intervals. We validate this approach empirically on two influenza-related Twitter datasets used for public health monitoring.

Background: Quantification

The quantification problem was first described in seminal work by Forman (2005; 2008), who showed that classification errors introduce systematic bias into the calculation of the number of positives. He used the term "classify and count" to describe the naïve quantification approach of simply counting the number of positively classified instances, and proposed several methods for adjusting the counts based on the true and false positive rates of the classifier, with some methods motivated specifically for data with imbalanced classes (Forman 2008).

This line of work has been extended to consider the effect of concept drift on quantification (Xue and Weiss 2009; Pérez-Gállego, Quevedo, and del Coz 2017), to count ordinal values (Da San Martino, Gao, and Sebastiani 2016), and to incorporate classifier probabilities into quantification estimates (Bella et al. 2010). See González et al. (2017) for a review of quantification methods.

In practice, quantification is an increasingly widespread application of social media posts. All of the health studies cited above in the introduction used the "classify and count" method of quantification (Forman 2008), though they did not refer to it as such; indeed, most work on aggregating social media content does not reference related work on quantification, even though quantification is implicitly being performed. After reviewing all papers on Google Scholar that cited the quantification papers above, we were able to find only a small number of studies that used adjustments

[*] LA-UR-18-24425

when quantifying social media posts, all for the application of sentiment analysis (Gao and Sebastiani 2015; 2016; Nakov et al. 2016; Sebastiani 2018). As far as we were able to discover, no work on social media-based health monitoring has applied adjustments when aggregating data.

Confidence Intervals

All previously proposed quantification methods have focused on producing point estimates of counts. We argue that for many quantification tasks it is useful to provide confidence intervals around the estimate; indeed, many of the social media studies we cited in the introduction constructed confidence intervals or similar statistics, but did not adjust for classification error. The main contribution of this work is to present an adjusted method for constructing bootstrap-based confidence intervals to correctly account for classification error, described in the next section. In our experiments, we show that naïvely-constructed confidence intervals are highly inaccurate, and our proposed algorithm is much more accurate than simply constructing confidence intervals using statistics adjusted with Forman's methods.

Adjusted Confidence Intervals

In this section, we present a non-parametric approach to constructing a confidence interval for the percent of instances within a group (e.g., the percent of tweets within a week) that are labeled positive. We denote this estimate as $\hat{p}$.

We first review bootstrapping for constructing confidence intervals, then propose a modification that incorporates classifier error into the sampling procedure.

Bootstrapping

Bootstrapping, or bootstrap resampling, is a procedure to simulate the statistics one would obtain when sampling from a distribution (Efron and Tibshirani 1993). A bootstrapped estimate is obtained by sampling N instances with replacement from the original dataset of size N, then calculating the statistic (e.g., $\hat{p}$) on the set of sampled instances. This procedure can be repeated many times to obtain many bootstrapped estimates, providing a distribution over estimates. To construct a $c\%$ confidence interval, the bootstrapped estimates can be sorted, and the range of the middle $c\%$ of values can be taken as the interval.

Error-adjusted Bootstrapping

If bootstrapping is applied to noisy classifications rather than true labels, then the samples will not be drawn from the correct distribution. We propose an adjustment to the sampling procedure that draws from the actual distribution of the data.

For each bootstrap sample, after selecting the instances (sampled with replacement), we randomly sample the labels of the instances two ways. The first is according to the confusion matrix of the classifier. If an instance is classified positive, we sample the label according to $P(Y_i|\hat{Y}_i = 1)$, where Y_i is the true label of instance i and $\hat{Y}_i$ is the classifier estimate. If an instance is classified negative, we sample the label according to $P(Y_i|\hat{Y}_i = 0)$. In this way, rather than treating the classifications as labels directly, we sample labels based on the probability that the classifier predicted an incorrect label. This procedure simulates the classification process in addition to the sampling process when obtaining an estimate.

Algorithm 1: Error-adjusted bootstrap resampling

```
Data: Set of $N$ instances classified as $\hat{Y}_i \in \{0,1\}$
Input: Number of bootstrap samples, $T$
Output: $S$, a set of $T$ estimates of $\hat{p}$
$S = \{\}$
for $1 \leq t \leq T$ do
    $\mathbf{y} = []$
    for $1 \leq i \leq N$ do
        Sample instance $j \in \{1, 2, \ldots, N\}$;
        if $\hat{Y}_j = 1$ then
            Sample $y \sim P(Y_j = y|\hat{Y}_j = 1)$;
        else
            Sample $y \sim P(Y_j = y|\hat{Y}_j = 0)$;
        end
        $\mathbf{y}$ += $[y]$;
    end
    $\hat{p} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_i$;
    $S = S \cup \{\hat{p}\}$;
end
return $S$
```

We refer to this approach as *error-adjusted* bootstrapping. The steps to obtain a set of error-adjusted bootstrapped samples are detailed in Algorithm 1.

Correctness of Algorithm The underlying assumption of bootstrap resampling is that the instances are i.i.d. and that uniformly sampling an instance is a draw from $P(Y)$. If the distribution of classifications $P(\hat{Y})$ is different from the distribution of labels $P(Y)$, then randomly sampling from the classifier outputs will not correctly draw from $P(Y)$.

Our approach uses the distribution $P(\hat{Y})$ and predictive values $P(Y|\hat{Y})$ to correctly calculate $P(Y)$: $P(Y_i = y) = P(Y_i = y|\hat{Y}_i = 0)P(\hat{Y}_i = 0) + P(Y_i = y|\hat{Y}_i = 1)P(\hat{Y}_i = 1)$.

As a generative process, sampling from this marginal distribution corresponds to the following steps for each instance i: (i) Sample $\hat{y}_i \sim P(\hat{Y})$; (ii) Sample $y_i \sim P(Y|\hat{Y}_i = \hat{y}_i)$. This matches Algorithm 1, which thus samples a label y according to to the true label distribution $P(Y)$ rather than the classification distribution $P(\hat{Y})$.

Predictive Value Estimation As described so far, we assume the positive predictive value, $P(Y|\hat{Y} = 1)$, and negative predictive value, $P(Y|\hat{Y} = 0)$, are known. We propose two approaches to estimating these values. The first uses cross-validation to provide point estimates of the positive and negative predictive values at each threshold of interest. This is the same approach used in prior work (Forman 2008).

The second approach extends Algorithm 1 to use a posterior distribution over predictive values. We do this by fitting a beta distribution to the individual estimates from cross-validation. We then draw a new estimate of the predictive values before sampling each label y_j during bootstrapping. We refer to this in experiments as the **extended** algorithm. Importantly, data used for these methods may be subject to other types of bias, including concept drift. If error rates change, predictive values would need to be re-estimated with new data (Pérez-Gállego, Quevedo, and del Coz 2017).

Experiments

We now experiment with estimating the percent of positive tweets in two datasets, comparing four different methods of constructing bootstrap-based confidence intervals.

Datasets and Classification Details

We experimented with binary classification on two datasets:

- **Flu Vaccination:** A set of 10,000 tweets labeled with if the tweet indicates that someone has received an influenza vaccination (i.e., a seasonal flu shot) (Huang et al. 2017) from 2013-2016. The aggregation task is to calculate the percent of tweets that indicate vaccination each month.
- **Flu Infection:** A set of 1,017 tweets from (Lamb, Paul, and Dredze 2013) from 2009 labeled as indicating flu infection. The original dataset included 5,000 tweets, but most are no longer available for download. The aggregation task is to calculate the percent of tweets indicating flu infection each week of available data.

Classification was done using binary logistic regression classifiers with unigram features implemented with `scikit-learn` (Pedregosa and others, 2011). For the larger *Flu Vaccination* data, we held out 15% of tweets for testing. Because the *Flu Infection* data were quite small, 25% of tweets were held out for testing. Grid search using five-fold cross validation on the training data was used to tune the ℓ_2 regularization parameter.

We experiment with different classification thresholds, meaning we set $\hat{y}_i = 1$ if $P(y_i = 1|x_i) > \tau$ for a threshold τ. Increasing the threshold will generally increase precision while reducing recall.

Baseline We experimentally compare to the "adjusted counts" method from Forman (2008). Here, the true positive rate (α) and the false positive rate (β) are used to obtain an adjusted estimate of the percent of positive instances:

$$p \approx \frac{\hat{p} - \beta}{\alpha - \beta}, \tag{1}$$

where $\hat{p}$ is the fraction estimated positive by the classifier. The estimate must be truncated to the range $[0, 1]$. In our experiments we calculate the adjusted counts within each bootstrapping iteration, and then construct confidence intervals of the adjusted counts.

Results

We examine the empirical characteristics of 95% confidence intervals constructed using bootstrap sampling, with and without making various error adjustments. We look at two characteristics: the fraction of times that the true value is contained in the interval (which should be 95%, asymptotically), as well as the size of the intervals.

Figure 1 shows these characteristics. The blue lines show the fraction of correct values contained in the 95% confidence intervals. As expected, the confidence intervals constructed using error-adjusted bootstrapping correctly capture the true values around 95% of the time, though it is less consistent on the smaller *Flu Infection* where the fraction sometimes drops to around 90%. This fraction is often higher than 95% with the extended version of Algorithm 1, suggesting that this method may unnecessarily overcompensate for uncertainty in the predictive values, but this method provides a benefit on the smaller *Flu Infection* set.

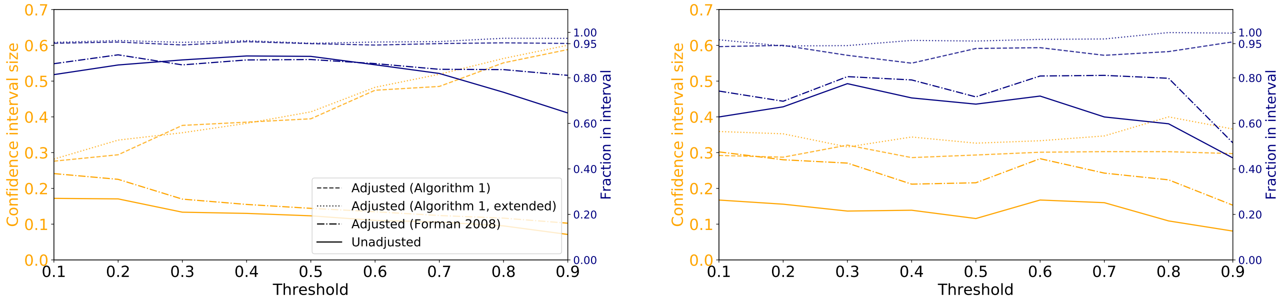
Importantly, we see that traditional bootstrapping without adjusting for classification error can severely affect the reliability of the confidence intervals. On *Flu Vaccination*, the unadjusted 95% confidence interval is correct less than 90% of the time at best and is as low as 65% at suboptimal thresholds. The Forman adjusted count method is more accurate than doing no adjustment, but is still inaccurate, with values between 80% and 90%. The situation is even worse on *Flu Infection*, where the unadjusted fraction is only 77% at best and as low as 45%. Similarly, the Forman baseline is more accurate than doing no adjustment, but less accurate than the Algorithm 1-adjusted methods, with a fraction around 80% at best.

Finally, the orange lines show the size of the intervals, to quantify how much wider the intervals must be to correctly adjust for error. In the *Flu Vaccination* dataset, the width of the confidence intervals in the Algorithm 1-adjusted methods consistently increase as the threshold increases even while the confidence intervals are consistently capturing the true values 95% of the time, suggesting that more statistical power can be obtained with a lower classification threshold (i.e., tuned for high recall). Due to the small size of the *Flu Infection* dataset, there is greater variation between the different methods, without clear conclusions.

Use Case: Vaccination Surveillance

Finally, we consider how this type of analysis relates to a real application of using the proportion of vaccine-related tweets to measure vaccination rates in a population. To do this, we applied the classifier trained on the Twitter dataset to a larger set of approximately 1 million tweets, from Huang et al. (2017). At different classification thresholds, we estimate the proportion of positive tweets in each month, and we compare these proportions to official flu vaccination data from the US Centers for Disease Control and Prevention (CDC), to evaluate how well monthly variations in vaccine tweets track true vaccination behavior (Huang et al. 2017). We measure this with Pearson correlation, calculating the proportions using adjusted bootstrapping from Algorithm 1 versus no adjustment.

Figure 2 shows the correlations between Twitter proportions and CDC data. While error-adjusted bootstrapping is more accurate at capturing confidence intervals (Figure 1), we do not see comparably large gains in correlations in this task. However, error-adjusted bootstrapping seems to provide a small benefit at some classification thresholds.



(a) Flu Vaccination

(b) Flu Infection

Figure 1: The size of 95% confidence intervals (orange) and fraction of true values contained within 95% confidence intervals (blue) at different classification thresholds, when constructing intervals with and without adjusting for error. With error-adjusted bootstrapping, the true value should theoretically be contained in the interval 95% of the time.

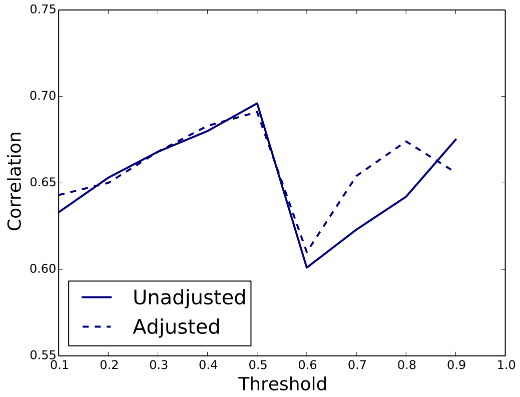


Figure 2: Correlations between Twitter classifier output and official vaccination data (higher is better).

Discussion and Conclusion

Confidence intervals constructed without accounting for classification error could be surprisingly inaccurate in our experiments (e.g., a 95% interval behaves like a 45% interval), highlighting the need to be careful about analyzing classifier outputs. We showed that a simple-to-implement adjustment to bootstrap sampling can correct for this, and we recommend this approach when aggregating social media posts or other filtered data.

References

Bella, A.; Ferri, C.; Hernandez-Orallo, J.; and Ramirez-Quintana, M. J. 2010. Quantification via probability estimators. In *ICDM*.

Culotta, A. 2010. Towards detecting influenza epidemics by analyzing Twitter messages. In *KDD*.

Da San Martino, G.; Gao, W.; and Sebastiani, F. 2016. Ordinal text quantification. In *SIGIR*.

Doan, S.; Ohno-Machado, L.; and Collier, N. 2012. Enhancing Twitter data analysis with simple semantic filtering: Example in tracking influenza-like illnesses. *arXiv preprint*.

Efron, B., and Tibshirani, R. J. 1993. *An Introduction to the Bootstrap*. Chapman & Hall.

Forman, G. 2005. Counting positives accurately despite inaccurate classification. In *ECML*.

Forman, G. 2008. Quantifying counts and costs via classification. *Data Min. Knowl. Discov.* 17(2):164–206.

Gao, W., and Sebastiani, F. 2015. Tweet sentiment: From classification to quantification. In *ASONAM*.

Gao, W., and Sebastiani, F. 2016. From classification to quantification in tweet sentiment analysis. *SNAM* 6(1):19.

González, P.; Castaño, A.; Chawla, N. V.; and Coz, J. J. D. 2017. A review on quantification learning. *ACM Comput. Surv.* 50(5):74:1–74:40.

Huang, X.; Michael C. Smith, M. J. P.; Ryzhkov, D.; Quinn, S. C.; Broniatowski, D. A.; and Dredze, M. 2017. Examining patterns of influenza vaccination in social media. In *AAAI Joint Workshop on Health Intelligence*.

Ji, X.; Chun, S. A.; and Geller, J. 2013. Monitoring public health concerns using twitter sentiment classifications. In *IEEE International Conference on Healthcare Informatics*.

Lamb, A.; Paul, M. J.; and Dredze, M. 2013. Separating fact from fear: Tracking flu infections on Twitter. In *NAACL*.

Mitra, T.; Counts, S.; and Pennebaker, J. 2016. Understanding anti-vaccination attitudes in social media. In *ICWSM*.

Nakov, P.; Ritter, A.; Rosenthal, S.; Sebastiani, F.; and Stoyanov, V. 2016. SemEval-2016 Task4: Sentiment analysis in Twitter. In *Proceedings of SemEval-2016*.

Paul, M. J., and Dredze, M. 2017. Social monitoring for public health. In *Synthesis Lectures on Information Concepts, Retrieval, and Services*. Morgan & Claypool. 1–185.

Pedregosa, F., and others,. 2011. Scikit-learn: Machine learning in Python. *JMLR* 12:2825–2830.

Pérez-Gállego, P.; Quevedo, J. R.; and del Coz, J. J. 2017. Using ensembles for problems with characterizable changes in data distribution: A case study on quantification. *Information Fusion* 34:87–100.

Sebastiani, F. 2018. Sentiment quantification of user-generated content. In *ESNAM*.

Xue, J. C., and Weiss, G. M. 2009. Quantification and semi-supervised classification methods for handling changes in class distribution. In *KDD*.