Analysis of Sampling Algorithms for Twitter

Deepan Palguna[†], Vikas Joshi[‡], Venkatesan Chakaravarthy[‡], Ravi Kothari[‡] and L V Subramaniam[‡]
[†]School of ECE, Purdue University, Indiana, USA
[‡]IBM India Research Lab, India
dpalguna@purdue.edu, {vijoshij, vechakra, rkothari, lvsubram}@in.ibm.com

Abstract

The daily volume of Tweets in Twitter is around 500 million, and the impact of this data on applications ranging from public safety, opinion mining, news broadcast, etc., is increasing day by day. Analyzing large volumes of Tweets for various applications would require techniques that scale well with the number of Tweets. In this work we come up with a theoretical formulation for sampling Twitter data. We introduce novel statistical metrics to quantify the statistical representativeness of the Tweet sample, and derive sufficient conditions on the number of samples needed for obtaining highly representative Tweet samples. These new statistical metrics quantify the representativeness or goodness of the sample in terms of frequent keyword identification and in terms of restoring public sentiments associated with these keywords. We use uniform random sampling with replacement as our algorithm, and sampling could serve as a first step before using other sophisticated summarization methods to generate summaries for human use. We show that experiments conducted on real Twitter data agree with our bounds. In these experiments, we also compare different kinds of random sampling algorithms. Our bounds are attractive since they do not depend on the total number of Tweets in the universe. Although our ideas and techniques are specific to Twitter, they could find applications in other areas as well.

1 Introduction

With Twitter's rising popularity, the number of Twitter users and Tweets are on a steady upward climb. It has been observed that around 500 million Tweets are generated daily. Twitter data is interesting and important since its impact on applications ranging from public safety, voicing sentiments about events [Mejova *et al.*, 2013], news broadcast, etc., is increasing day by day. The number of Tweets being so large means that there is a necessity to come up with strategies to deal with this data in a scalable fashion. One possible approach to deal with such large amounts of data is to sample the data and work with a subset of the original universe. Sampling Tweets has multiple uses. For example, extractive summarization of Tweets for human readers is essentially a sampling algorithm. Additionally, Twitter itself uses a 1% random sample in its API, which immediately raises questions such as: Can we produce representative samples using simple methods like random sampling? How many Tweets should we randomly sample so that the information contained in the universe is also contained in the sample? In this work, we propose different statistical properties to characterize the goodness of a random sample in the context of Twitter. We derive sufficient conditions on the number of samples needed in order to achieve the proposed goodness metrics. For theoretical analysis, we sample Tweets in a uniform random fashion with replacement.

Topic modeling in Twitter [Hong and Davison, 2010] is done using distributions over keywords which could include nouns like the names of personalities, places, events, etc. Twitter is also widely used to convey opinions or sentiments about these nouns. Motivated by this, we consider two statistical properties to quantify the representativeness of a sample. These are the frequency of occurrence of nouns and the conditional frequencies of sentiments, conditioned on each frequent noun. These frequencies constitute the conditional sentiment distribution of the noun. In order to compute this distribution, we assume that there are three possible sentiments for each Tweet—positive, negative or neutral. The conditional frequencies are computed as empirical frequencies as explained in later sections. In our theoretical analysis and experiments, we assume that a perfect parts of speech tagger [Bird *et al.*, 2009] provides the information on whether a word is a noun or not. Similarly, we assume that our sentiment analysis algorithm accurately classifies the sentiment of each Tweet into one of the three categories.

An overview of our contributions is presented now. We theoretically derive sufficient conditions on the number of samples needed so that:
a) nouns that are frequent in the original universe are frequent in the sample, and infrequent nouns in the universe are infrequent in the sample.
b) the conditional sentiment distributions of frequent nouns in the universe are close to the corresponding conditional sentiment distributions in the sample.
c) the dominant sentiment conditioned on a noun is also dom-

inant in the sample.

The exact statements of these results will be made in later sections. Our bounds are attractive since they do not depend on the number of Tweets in the universe, and depend only on the maximum length of a Tweet which can be readily bounded. We evaluate our theoretical bounds by performing experiments with real Twitter data.

To the best of our knowledge, ours is the first work dealing with ideas of frequent noun mining, which is analogous to frequent itemset mining in the database literature. It is also the first work to deal with ideas of restoring sentiment distributions associated with frequent nouns and preserving dominant sentiments associated with any particular noun. So the articulation and theoretical treatment of these ideas in sampling Twitter data distinguish our work from the current literature. Our ideas and results would be of interest to data providers and researchers in order to decide how much data to use in their respective applications. They could be used as a simple first step in extractive summarization or opinion mining in order to save on computation, especially when Tweets spanning multiple days are to be analyzed.

In the next section we give a literature review, following which, we set up the problem and introduce mathematical notations in Section 3. Sections 4 and 5 are used to state our theoretical results on frequent noun mining and sentiment restoration respectively. We then show our simulations and experimental results with real Twitter data in Section 6. All the main proofs are in the Appendix.

2 Related Work

Summarization algorithms for Twitter aim to extract representative Tweets from a large universe for human consumption. These algorithms are closely related to document summarization methods [Nenkova and McKeown, 2012]. Extractive Tweet summarization algorithms have been studied in the literature by [Sharifi *et al.*, 2010], [Yang *et al.*, 2012], [Chakrabarti and Punera., 2011]. A graph based summarization algorithm for a small universe size of 1000 has been proposed in [Liu *et al.*, 2012]. A comparison of various methods for summarizing Twitter data can be found in [Inouye and Kalita, 2011]. These existing Tweet summarization methods are often computationally expensive and would not scale to millions of Tweets, which is typical of the daily Tweet volume. Therefore, as mentioned previously, random sampling could serve as a first step before using any of these algorithms for human readable summaries, provided the random sample is representative of the universe.

There have been many papers that deal with the quality of samples given by Twitter API's free 1% sample. These papers are primarily empirical in nature. In [Morstatter *et al.*, 2013] the authors compare the API's feed with the Tweets obtained from Firehose—which is also provided by Twitter, but contains all the Tweets instead of a sample. In [Ghosh *et al.*, 2013], the authors empirically compare random sampling against sampling done with the help of human experts. In [Morstatter *et al.*, 2014], the authors analyze the bias in Twitter's API without using the costly Firehose data. The effects of using multiple streaming APIs is considered by [Joseph *et al.*, 2014]. As can be seen, there have been many papers on the empirical analysis of the quality of Tweets summaries and the Twitter API itself. It is also understood that random sampling preserves statistical properties of the universe. But to the best of our knowledge there has not been a treatment of the problem from our viewpoint of deriving conditions on sample sizes needed to produce representative samples, and ours is the first study to theoretically characterize the behaviour of random Tweet sampling.

3 Problem Formulation

The set of Tweets that form our universe of Tweets is denoted by T. This set contains N Tweets, each of which can have a maximum of L words. From N Tweets, we uniformly and randomly sample K Tweets with replacement. The set of Tweets in the sample is denoted using S. Each Tweet is modelled as a bag of words. Moreover, we assume that each Tweet has an associated sentiment, which can be in one of three possible classes—positive, negative or neutral. Some notations used throughout the paper:

$$K_w = \#_{\text{Tweets in } T \text{ that contain noun } w}$$
$$K_w^S = \#_{\text{Tweets in } S \text{ that contain noun } w}$$
$$K_{w,p} = \#_{\text{positive sentiment Tweets in } T \text{ that contain noun } w}$$
$$K_{w,p}^S = \#_{\text{positive sentiment Tweets in } S \text{ that contain noun } w}$$
$$f_w = \frac{K_w}{N},\ f_w^S = \frac{K_w^S}{K} \text{ (Noun frequency in universe \& sample)}$$
$$f_{w,p} = \frac{K_{w,p}}{K_w},\ f_{w,p}^S = \frac{K_{w,p}^S}{K_w^S} \text{ (Conditional sentiment frequency)}$$

Similar notations are used for minus (negative) and neutral sentiments, where we replace p (plus) in the subscript with m and n to denote minus and neutral respectively. The ordered triple of $(f_{w,p}, f_{w,m}, f_{w,n})$ forms the sentiment distribution conditional on noun w. As mentioned in the introduction, we derive sufficient conditions on the number of samples needed so that important statistical properties of the universe are retained in the sample with the desired probability. We now clarify two points related to our theoretical analysis.

Firstly, the sampling algorithm we use for analysis is a *batch* sampling method, which fixes a sample size and then does uniform random sampling with replacement. But when Twitter's API gives a 1% sample, it is very likely that it would perform *sequential* sampling, where each Tweet in the universe would be sampled one-by-one and independent of all other Tweets with probability 1%. We use batch sampling for analysis since the analysis is much cleaner than the case of sequential sampling. Experiments with real data suggest that the performance of both methods depend strongly only on the number of samples, and not on whether sampling is done in batch mode or in sequential mode.

Secondly, our sampling method is random and it may theoretically be possible to derive equations for the exact probabilities of the associated events as a function of K. But computing them and solving the resulting non-linear equations for the sufficient K would be very difficult. Doing this might give stronger bounds for K, but its infeasibility motivates us to derive sufficient conditions on K using bounds on the probabilities instead of the exact probabilities. The sufficient

sample sizes would depend only on parameters that govern probabilistic goodness guarantees demanded by the application and not on N, which is an added benefit. In the next two sections, we describe the statistical properties that we want to retain, the associated probabilistic goodness guarantees, and state our results on the sample size bounds.

4 Sampling for $\theta-$frequent Nouns

For a number $\theta \in [0,1]$, a noun whose frequency in T is at least θ is said to be a $\theta-$frequent noun. In this section, we derive a sufficient condition on the number of Tweets to be sampled in order to accomplish two goals. For some $\epsilon \in [0,1]$, these two goals are:

- Nouns that occur with frequency more than θ in T occur with frequency more than $(1-\epsilon/2)\theta$ in S.
- Nouns that occur with frequency lesser than $(1-\epsilon)\theta$ in T occur with frequency less than $(1-\epsilon/2)\theta$ in S.

We declare a failure event to have occurred if after sampling K Tweets, the algorithm fails to accomplish any one of these two goals. This idea and the theoretical results that follow are very closely related to $\theta-$frequent itemset mining in the database literature. Following the analysis in [Chakaravarthy *et al.*, 2009], if noun w is $\theta-$frequent, then

$$\mathbb{P}\{f_w^S \leq (1-\epsilon/2)\theta\} \leq \exp\{-\epsilon^2 K\theta/8\}.$$

If noun w's frequency is lesser than $(1-\epsilon)\theta$, we get

$$\mathbb{P}\{f_w^S \geq (1-\epsilon/2)\theta\} \leq \exp\{-\epsilon^2(1-\epsilon)K\theta/12\}.$$

To bound the failure probability, we have to bound the probability of failure over all nouns. We can accomplish this using the union bound in conjunction with a bound on the number of nouns. If the dictionary contains M nouns, then the number of nouns in the Tweet universe is upper bounded by $\min\{M, NL\}$. So we will get that the failure probability is upper bounded by h if

$$K \geq \frac{12}{\epsilon^2(1-\epsilon)\theta}(\min\{\log(M), \log(NL)\} - \log(h))$$

This bound depends on the number of nouns and on the size of the Tweet universe, whereas we would like to derive a bound that depends only on L, since L is easier to upper bound than M or N. In order to derive this bound, we use the results of [Chakaravarthy *et al.*, 2009]. In this direction, we first derive a simple upper bound for the number of $\theta-$frequent nouns.

Lemma 1. *The number of nouns that occur with frequency at least θ is upper bounded by L/θ.*

Proof. The total number of words $\leq NL$. A θ-frequent noun would consume at least $N\theta$ words. So the number of θ-frequent nouns cannot exceed L/θ. □

Lemma 2. *If the size of the sample obeys $K \geq \frac{24}{\epsilon^2(1-\epsilon)\theta}(\log\frac{8L}{(1-\epsilon)\theta h} + 5)$, then the probability of a failure is upper bounded by $4h$ for $h < 1/4$.*

Proof. We omit the full proof here since it follows from the proof of Theorem 3.1 in [Chakaravarthy *et al.*, 2009] and from Lemma 1. □

5 Sampling for Sentiment Restoration

Since Twitter is used as an important medium for voicing opinions, it is important for the conditional sentiment distribution of nouns in the summary to be close to the corresponding distribution in the universe. For some $\lambda \in [0,1]$, we say that the sentiment distribution of word w is *not restored* if the following event occurs

$$\{|f_{w,p}^S - f_{w,p}| > \lambda\} \cup \{|f_{w,m}^S - f_{w,m}| > \lambda\} \cup \{|f_{w,n}^S - f_{w,n}| > \lambda\} \quad (1)$$

We say that a failure occurs if even a single $\theta-$frequent noun's sentiment distribution is not restored. We want to sample enough Tweets so that the probability of failure is bounded above by $h \in [0,1]$.

Theorem 1. *If $K \geq \dfrac{\log((h\theta)/(6L))}{\log(1-\theta+\theta\exp\{-\lambda^2/3\})}$, then the probability of failure to restore sentiment is less than h.*

Proof. Proof is in the Appendix 8.1. □

In opinion mining, one may not want to guarantee that the entire distribution conditioned on noun w be restored. One may rather be interested to simply guarantee that the dominant sentiment in the universe also dominates in the sample. A sentiment is said to dominate if its frequency conditioned on w is higher than the other two conditional frequencies. For example, if noun w is such that $f_{w,p} > f_{w,m}$ and $f_{w,p} > f_{w,n}$, then positive sentiment is said to be the dominant sentiment in the universe. The goal of dominant sentiment preservation is to sample enough Tweets so that this property is *preserved* with high probability in the sample as well i.e., $f_{w,p}^S > f_{w,m}^S$ and $f_{w,p}^S > f_{w,n}^S$. For deriving the sufficient number of Tweets to guarantee this, we develop a preliminary result in Lemma 3. Let (X_1, X_2, X_3) be jointly multinomial random variables with parameters (M, f_1, f_2, f_3), where $f_1 + f_2 + f_3 = 1$ and $f_1 > f_2 > f_3$. Consider first

$$\begin{aligned}\mathbb{P}\{X_3 > X_1\} &= \mathbb{E}[\mathbb{P}\{X_3 > X_1 \mid X_2 = x_2\}] \\ &= \mathbb{E}[\mathbb{P}\{X_3 > (M-x_2)/2 \mid X_2 = x_2\}] \end{aligned} \quad (2)$$

It can be shown that conditional on $X_2 = x_2$, X_3 is binomial with parameters $(M - x_2, f_3' = f_3/(f_1+f_3))$.

Lemma 3. *For $\delta_3 = \frac{1}{2f_3'} - 1$, $\mathbb{P}\{(X_3 > X_1) \cup (X_2 > X_1)\}$ $\leq 2[f_2 + (1-f_2)\exp\{-(\delta_3^2 f_3')/(2+\delta_3)\}]^M$*

Proof. Proof is in the Appendix 8.2. □

We can use Lemma 3 with similar notations to derive a bound on the number of Tweets that are adequate to *preserve* the dominant sentiment for noun w. We relabel $f_{w,p}$, $f_{w,m}$ and $f_{w,n}$ using $f_{w,1}$, $f_{w,2}$ and $f_{w,3}$ such that they are ordered as $f_{w,1} > f_{w,2} > f_{w,3}$. The same notation is followed for $K_{w,1}$, $K_{w,2}$ and $K_{w,3}$. Similarly, we use $f_{w,3}' = \frac{f_{w,3}}{f_{w,1}+f_{w,3}}$ and $\delta_{w,3} = \frac{1}{2f_{w,3}'} - 1$.

Theorem 2. *For $h \in [0,1]$, the probability that noun w's dominant sentiment is not dominant in the sample is upper bounded by h if $K \geq$*

$$\frac{\log(h/2)}{\log\left(1 - f_w + f_w\big(f_{w,2} + (1-f_{w,2})\exp\big\{-\frac{\delta_{w,3}^2}{2+\delta_{w,3}} f_{w,3}'\big\}\big)\right)}$$

Proof. Proof is in Appendix 8.3. □

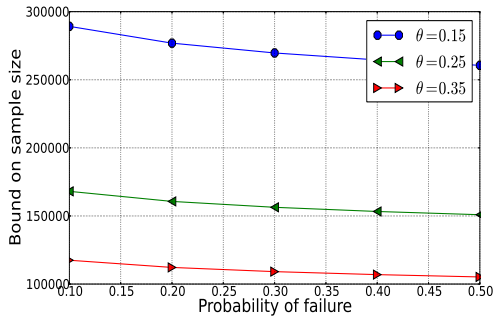

Figure 1: Bounds on sample size for θ-frequent noun mining for $\epsilon = 0.1$ and as a function of the failure probability.

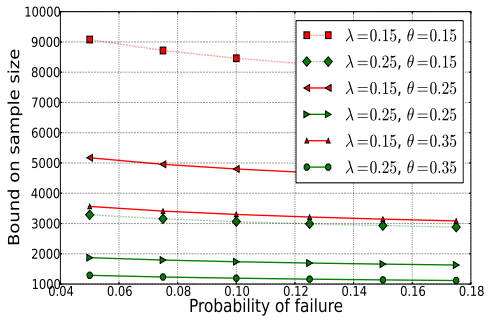

Figure 2: Bounds for sentiment restoration over all θ-frequent nouns as a function of the failure probability.

6 Simulations and Experiments

6.1 Data Set Description

Using Twitter's free API, we created a data set based on certain important political events of 2014, which we call the PE-2014 data set. This is done by using a comprehensive vocabulary to filter and retain only those Tweets that pertain to certain political events. These Tweets form our Tweet universe with a volume of $\sim 200,000$ Tweets per day. Although this is substantially smaller than the total daily Tweet volume, it could still be too large a set for more sophisticated summarization algorithms, especially if we want to summarize Tweets from multiple days. Our simulation results are obtained over one day's worth of data (March 22, 2014).

6.2 Theoretical Bounds

To characterize the bounds given by our theoretical results, we need an estimate or an upper bound for L. For these results, we estimate L to be 33 words by going through the entire PE-2014 data set once. If going through the universal Tweet set once is too costly, then an upper bound for L would be 140, which is the maximum number of characters allowed per Tweet. Fig. 1 shows the number of samples needed to achieve the conditions in Lemma 2. While some of the sample sizes could be large, the values are still smaller and independent of the total daily Tweet volume. Fig. 2 shows the bound in Theorem 1 as a function of the probability of failure to restore the sentiment distribution of all $\theta-$frequent nouns. This is plotted for different values of the deviation λ and θ. The values in Fig. 2 are much smaller that the values in Fig. 1. This is because in Theorem 1, we do not impose any conditions on the frequency of $\theta-$frequent nouns in the sample. Therefore for these settings of θ, ϵ and λ, the sufficient condition on sample size for frequent noun mining is also sufficient for sentiment preservation. In [Riondato and Upfal, 2012], the authors derive bounds for frequent itemset mining in terms of the VC dimension. This could give us tighter bounds for frequent noun mining, but it is part of our future work.

The leftmost panel of Fig. 3 compares the theoretical bounds for dominant sentiment preservation (Theorem 2) and sentiment distribution restoration for a single noun. For the same failure probability, the bounds on the sample size for preserving the dominant sentiment are much smaller than the bound on the sample size for restoring the entire distribution. In the middle panel of Fig. 3, we compare the two cases of dominant sentiment preservation shown in the left panel. One would expect the case with $f_{w,1} = 0.9, f_{w,2} = 0.06, f_{w,3} = 0.04$ to have a smaller sample size bound as compared to $f_{w,1} = 0.45, f_{w,2} = 0.35, f_{w,3} = 0.2$, due to the closeness of frequencies in the latter case. This intuition is confirmed in the middle panel. It is also possible to encounter cases when the bounds for order preservation are larger than the bounds for distribution restoration as seen in the rightmost panel of Fig. 3. Since the conditional frequencies of the sentiments are very close to each other ($f_{w,1} = 0.35, f_{w,2} = 0.33, f_{w,3} = 0.32$), the number of samples needed to preserve their order exceeds the number of samples needed to restore their distribution for these parameters. In such cases though, preserving sentiment order may not be important since human users may conclude that no sentiment really dominates.

Table 1: Errors in $\theta-$frequent noun mining in PE-2014, shown for $\theta = 0.1$, $\epsilon = 0.1$ and 100 Monte Carlo rounds

# samples (K)	# freq. noun errors	Mean max. deviation
2000	7	0.0952
4000	3	0.0706
8000	0	0.0488
10000	0	0.0426

6.3 Experiments with Real Data

We measure the sample quality in terms of statistical properties and failure events related to the theoretical formulation, and compare them with the theoretical results for a range of sample sizes. For identifying $\theta-$frequent nouns, the error events measured over 100 Monte Carlo rounds are shown in the second column of Table 1. We measure the maximum deviation in sentiment for word w as

$$\max\{|f^S_{w,p} - f_{w,p}|, |f^S_{w,m} - f_{w,m}|, |f^S_{w,n} - f_{w,n}|\} \qquad (3)$$

The third column of Table 1 shows the maximum deviation in Eq. 3, maximized over the $\theta-$frequent nouns and averaged over 100 Monte Carlo rounds.

From the second column of Table 1, we infer that for $\theta = 0.1$ and $\epsilon = 0.1$ it would be sufficient to sample around

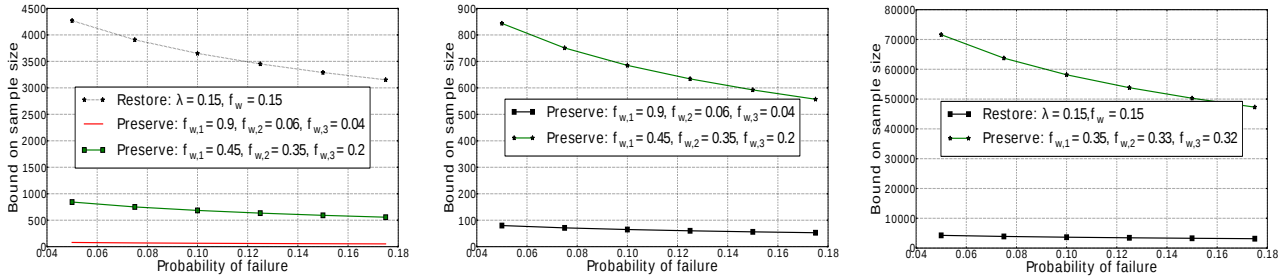

Figure 3: Sentiment distribution restoration vs. dominance preservation for one noun with $f_w = 0.15$, allowed deviation $\lambda = 0.15$.

Table 2: Error events in preserving the dominant sentiment for PE-2014 data, shown for three most frequent nouns and 100 Monte Carlo rounds. The fourth noun is an illustration of a case with difficult order preservation. Two different sample sizes and their corresponding error counts are shown as ordered tuples for each noun.

Noun number	Noun frequency (f_w)	Sentiments $(f_{w,p}, f_{w,m}, f_{w,n})$	Number of samples	Errors in order restoration
Noun 1	0.25	(0.31, 0.25, 0.44)	(2000, 4000)	(1, 0)
Noun 2	0.14	(0.32, 0.19, 0.49)	(2000, 4000)	(0, 0)
Noun 3	0.12	(0.23, 0.27, 0.50)	(2000, 4000)	(0, 0)
Noun 4	0.054	(0.378, 0.255, 0.367)	(2000, 4000)	(39, 38)

8000 Tweets in order to satisfy the $\theta-$frequent noun mining condition for these parameters. In Lemma 2, the ϵ parameter, the constant factor, and the need to make infrequent nouns of the universe infrequent in the sample makes the theoretical bounds in Fig. 1 much larger than the ones needed in practise. Also, we see from the third column of Table 1 that it is enough to sample 2000 Tweets to achieve a mean maximum sentiment deviation of 0.0952 for $\theta = 0.1$. These agree with the sufficient conditions in Theorem 1 and are quite close to the bounds shown in Fig 2. Table 2 shows that the number of errors in sentiment order preservation for three most frequent nouns is negligible. The last line is for a fourth noun whose sentiment order is intuitively hard to preserve.

6.4 Comparison with Sequential Sampling

In sequential sampling we go through the Tweets in the universe one-by-one and sample a Tweet independent of others with probability p. For the same data set, we perform sequential sampling with two different p values—0.01 and 0.02. For each value of p, we perform 100 Monte Carlo rounds and measure the same metrics as in batch sampling. These results are shown in Table 3. From this table, we see that the performance of sequential sampling and random sampling with replacement (shown in Table 1) are very similar, and are primarily influenced by the sample size alone.

7 Conclusion

In this work we have derived bounds on the number of random samples needed to guarantee statistically representative Tweet samples. Random sampling can be used as a first step before using other sophisticated algorithms to form human readable summaries or for mining social opinion. For example, if we want to find out Twitter sentiment for a keyword that we know is frequent in the universe, then we could run our sentime nt analysis tools on a much smaller random sample whose size does not depend on N. Therefore, using random sampling as a first step could provide significant computational savings for many such applications with practical accuracy requirements. Our experiments with real data confirm that the bounds agree with the sample sizes needed in practice. Our results would readily apply to frequent topic mining and restoring topic models represented as mixture distributions in order to sample a diverse set of Tweets. These ideas also easily extend to applications beyond Twitter to other short messages such as call center transcripts. The theoretical analysis of sequential sampling is part of our future work. Moreover, the bounds for frequent noun mining could be made tighter by using the VC dimension ideas from [Riondato and Upfal, 2012], which is also part of our future work.

8 Appendix: Proofs

8.1 Proof of Theorem 1

The event of the sentiment distribution of word w not being preserved is:

$$\{|f^S_{w,p} - f_{w,p}| > \lambda\} \cup \{|f^S_{w,m} - f_{w,m}| > \lambda\} \cup \{|f^S_{w,n} - f_{w,n}| > \lambda\}$$

First consider the event $\{|f^S_{w,p} - f_{w,p}| > \lambda\}$ conditional on K^S_w. Conditional on a realization of K^S_w, $K^S_{w,p}$ is a binomial random variable with parameters K^S_w and $f_{w,p}$. So we can use Chernoff's bound [Chernoff, 1952] for binomial random vari-

Table 3: Error events in $\theta-$frequent noun mining in PE-2014 data shown for $\theta = 0.1$, $\epsilon = 0.1$ and 100 Monte Carlo rounds. The number of sentiment order preservation errors in the last column is the sum over the 3 most frequent nouns shown in Table 2.

p	$\lfloor$Avg. # of samples$\rfloor$	# error events in $\theta-$ freq. noun mining	Mean max. deviation	# sentiment order preservation errors
0.01	2385	0	0.0883	0
0.02	4790	0	0.0627	0

ables to derive an upper bound on the probability of the event.

$$\mathbb{P}\left\{|f^S_{w,} - f_{w,p}| > \lambda \middle| K^S_w\right\} =$$
$$\mathbb{P}\left\{K^S_{w,p} > K^S_w \frac{K_{w,p}}{K_w}\left(1 + \lambda \frac{K_w}{K_{w,p}}\right)\middle| K^S_w\right\}$$
$$+ \mathbb{P}\left\{K^S_{w,p} < K^S_w \frac{K_{w,p}}{K_w}\left(1 - \lambda \frac{K_w}{K_{w,p}}\right)\middle| K^S_w\right\}$$
$$\leq \exp(-\frac{\lambda^2 K^S_w}{3f_{w,p}}) + \exp(-\frac{\lambda^2 K^S_w}{2f_{w,p}}) \leq 2\exp(-\frac{\lambda^2 K^S_w}{3f_{w,p}}) \quad (4)$$

We can remove the conditioning by taking expectation over all possible realizations of K^S_w to get:

$$\mathbb{P}\{|f^S_{w,p} - f_{w,p}| > \lambda\} \leq \mathbb{E}\left[2\exp(-\frac{\lambda^2 K^S_w}{3f_{w,p}})\right] \quad (5)$$

The expectation in the Eq. 5 is over the realizations of K^S_w. This expectation is the moment generating function of a binomial random variable with parameters (K, f_w) evaluated at $-\lambda^2/(3f_{w,p})$ [Papoulis and Pillai, 2002]. Using the union bound and adding the bounds for the three sentiments we have

$$\mathbb{P}\{\text{Sentiment not preserved for noun } w\}$$
$$\leq 6[1 - f_w + f_w \exp(-\lambda^2/(3\max\{f_{w,p}, f_{w,m}, f_{w,n}\}))]^K$$
$$\leq 6[1 - f_w + f_w \exp(-\lambda^2/3)]^K \quad (6)$$

In order to bound the probability of failure, which is the probability that the sentiment distribution of at east one $\theta-$frequent noun is not preserved, we again apply the union bound over all $\theta-$frequent nouns. The function of f_w on the R.H.S. of Eq. 6 decreases with f_w. Since we have taken w to be $\theta-$frequent, this expression is maximized by substituting θ in place of f_w. Moreover, the bound on the number of $\theta-$frequent nouns is L/θ, which gives the upper bound on the probability of failure to preserve sentiment for any θ frequent noun as $(6L/\theta)[1 - \theta + \theta\exp(-\lambda^2/3)]^K$. So if $K \geq \frac{\log((h\theta)/(6L))}{\log(1 - \theta + \theta\exp\{-\lambda^2/3\})}$ then the probability of failure will not exceed h—proving Theorem 1.

8.2 Proof of Lemma 3

$$\mathbb{P}\big\{X_3 > (M - x_{2/}(2)\big|X_2 = x_2\big\}$$
$$= \mathbb{P}\left\{X_3 > (M - x_2)f'_3(\frac{1}{2f'_3} - 1 + 1)\middle| X_2 = x_2\right\}$$
$$\leq \exp\big\{-(\delta^2_3(M - x_2)f'_3)/(2 + \delta_3)\big\} \quad (7)$$

The application of Chernoff's bound to get Eq. 7 is possible since $f_1 > f_3$ resulting in $f'_3 < 1/2$. The marginal distribution of $M - X_2$ is binomial with parameters $(M, 1 - f_2)$. Taking expectation w.r.t. X_2, we get

$$\mathbb{P}\{X_3 > X_1\} \leq [f_2 + (1 - f_2)\exp\{-(\delta^2_3 f'_3)/(2 + \delta_3)\}]^M$$

By symmetry

$$\mathbb{P}\{X_2 > X_1\} \leq [f_3 + (1 - f_3)\exp\{-(\delta^2_2 f'_2)/(2 + \delta_2)\}]^M$$

Now we can use the union bound and write the probability that X_1 is smaller than either X_2 or X_3 as

$$\mathbb{P}\{(X_3 > X_1) \cup (X_2 > X_1)\}$$
$$\leq [f_3 + (1 - f_3)\exp\{-(\delta^2_2 f'_2)/(2 + \delta_2)\}]^M$$
$$+ [f_2 + (1 - f_2)\exp\{-(\delta^2_3 f'_3)/(2 + \delta_3)\}]^M \quad (8)$$

The bound in Eq. 8 can be further simplified by considering $\frac{\delta^2_3}{2+\delta_3} f'_3$. This is equal to $\frac{f'_3}{(2f'_3-1)(4f'_3+1)}$, which is a decreasing function of f'_3 if $f'_3 < 1/2$. This holds because we have taken f'_3 to be $\frac{f_3}{f_1+f_3}$ and assumed that $f_1 > f_3$. Since $f_2 > f_3$, we have $f'_2 = \frac{f_2}{f_1+f_2} > \frac{f_3}{f_1+f_3} = f'_3$. Consequently, $\exp\{-\frac{\delta^2_3}{2+\delta_3} f'_3\} > \exp\{-\frac{\delta^2_2}{2+\delta_2} f'_2\}$. Now we can write the bound in Eq. 8 as

$$\mathbb{P}\{(X_3 > X_1) \cup (X_2 > X_1)\}$$
$$\leq 2[f_2 + (1 - f_2)\exp\{-(\delta^2_3 f'_3)/(2 + \delta_3)\}]^M \quad (9)$$

This concludes the proof of Lemma 3.

8.3 Proof of Theorem 2

We follow a similar proof by conditioning first on a realization of K_w. Conditioned on K_w, $(K_{w,1}, K_{w,2}, K_{w,3})$ are jointly multinomial with parameters $(K_w, f_{w,1}, f_{w,2}, f_{w,3})$. From Lemma 3, we have

$$\mathbb{P}\{\text{Noun } w\text{'s sentiment not preserved} \mid K_w\} \leq$$
$$2[f_{w,2} + (1 - f_{w,2})\exp\{-(\delta^2_{w,3} f'_{w,3})/(2 + \delta_{w,3})\}]^{K_w} \quad (10)$$

We average over K_w again whose marginal distribution is binomial with parameters (K, f_w). This gives the probability generating function [Papoulis and Pillai, 2002].

$$\mathbb{P}\{\text{Noun } w\text{'s sentiment not preserved}\} \leq$$
$$2\left[1 - f_{w,2} + f_w(f_{w,2} + (1 - f_{w,2})\exp\left\{-\frac{\delta^2_{w,3}}{2 + \delta_{w,3}} f'_{w,3}\right\}\right]^K$$

So if

$$K \geq \frac{\log(h/2)}{\log\left(1 - f_w + f_w\big(f_{w,2} + (1 - f_{w,2})\exp\big\{-\frac{\delta^2_{w,3}}{2+\delta_{w,3}} f'_{w,3}\big\}\big)\right)}$$

then the probability of noun w's sentiment is not preserved is lesser than h. This proves Theorem 2.

References

[Bird *et al.*, 2009] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition, 2009.

[Chakaravarthy *et al.*, 2009] Venkatesan T Chakaravarthy, Vinayaka Pandit, and Yogish Sabharwal. Analysis of sampling techniques for association rule mining. In *Proceedings of the 12th international conference on database theory*, pages 276–283. ACM, 2009.

[Chakrabarti and Punera., 2011] D. Chakrabarti and K. Punera. Event summarization using tweets. In *ICWSM*, 2011.

[Chernoff, 1952] Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507, 12 1952.

[Ghosh *et al.*, 2013] Saptarshi Ghosh, Muhammad Bilal Zafar, Parantapa Bhattacharya, Naveen Sharma, Niloy Ganguly, and Krishna Gummadi. On sampling the wisdom of crowds: Random vs. expert sampling of the twitter stream. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1739–1744. ACM, 2013.

[Hong and Davison, 2010] Liangjie Hong and Brian D Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, pages 80–88. ACM, 2010.

[Inouye and Kalita, 2011] David Inouye and Jugal K Kalita. Comparing twitter summarization algorithms for multiple post summaries. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 298–306. IEEE, 2011.

[Joseph *et al.*, 2014] Kenneth Joseph, Peter M Landwehr, and Kathleen M Carley. Two 1% s dont make a whole: Comparing simultaneous samples from twitters streaming api. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 75–83. Springer, 2014.

[Liu *et al.*, 2012] Xiaohua Liu, Yitong Li, Furu Wei, and Ming Zhou. Graph-based multi-tweet summarization using social signals. In *COLING'12*, pages 1699–1714, 2012.

[Mejova *et al.*, 2013] Yelena Mejova, Padmini Srinivasan, and Bob Boynton. Gop primary season on twitter: "popular" political sentiment in social media. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 517–526, New York, NY, USA, 2013. ACM.

[Morstatter *et al.*, 2013] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M. Carley. Is the sample good enough? comparing data from Twitter's streaming API with Twitter's Firehose. *Proceedings of ICWSM*, 2013.

[Morstatter *et al.*, 2014] Fred Morstatter, Jürgen Pfeffer, and Huan Liu. When is it biased?: Assessing the representativeness of twitter's streaming api. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, pages 555–556, 2014.

[Nenkova and McKeown, 2012] Ani Nenkova and Kathleen McKeown. A survey of text summarization techniques. In *Mining Text Data*, pages 43–76. Springer US, 2012.

[Papoulis and Pillai, 2002] A. Papoulis and S.U. Pillai. *Probability, random variables, and stochastic processes*. McGraw-Hill electrical and electronic engineering series. McGraw-Hill, 2002.

[Riondato and Upfal, 2012] Matteo Riondato and Eli Upfal. Efficient discovery of association rules and frequent itemsets through sampling with tight performance guarantees. In *Machine Learning and Knowledge Discovery in Databases*, pages 25–41. 2012.

[Sharifi *et al.*, 2010] Beaux Sharifi, Mark-Anthony Hutton, and Jugal K. Kalita. Experiments in microblog summarization. In *Proceedings of the 2010 IEEE Second International Conference on Social Computing*, SOCIALCOM '10, pages 49–56, Washington, DC, USA, 2010. IEEE Computer Society.

[Yang *et al.*, 2012] Xintian Yang, Amol Ghoting, Yiye Ruan, and Srinivasan Parthasarathy. A framework for summarizing and analyzing twitter feeds. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 370–378, New York, NY, USA, 2012. ACM.