# NAVAL POSTGRADUATE SCHOOL

**MONTEREY, CALIFORNIA**

# THESIS

**AUTHOR DETECTION ON A MOBILE PHONE**

by

Jody Grady

March 25, 2011

| | |
|---|---|
| Thesis Advisor: | Dr. Rob Beverly |
| Second Reader: | Dr. Craig Martell |

**Approved for public release; distribution is unlimited**

THIS PAGE INTENTIONALLY LEFT BLANK

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 0704–0188*



| 1. REPORT DATE *(DD–MM–YYYY)* | 2. REPORT TYPE | 3. DATES COVERED *(From — To)* |
|---|---|---|
| 18–3–2011 | Master's Thesis | 2009-03-01 - 2011-03-25 |

**4. TITLE AND SUBTITLE**

Author Detection on a Mobile Phone

**5a. CONTRACT NUMBER**

**5b. GRANT NUMBER**

**5c. PROGRAM ELEMENT NUMBER**

**6. AUTHOR(S)**

Jody Grady

**5d. PROJECT NUMBER**

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**

Naval Postgraduate School
Monterey, CA 93943

**8. PERFORMING ORGANIZATION REPORT NUMBER**

**9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)**

Department of the Navy

**10. SPONSOR/MONITOR'S ACRONYM(S)**

**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

**12. DISTRIBUTION / AVAILABILITY STATEMENT**

Approved for public release; distribution is unlimited

**13. SUPPLEMENTARY NOTES**

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. **IRB Protocol Number: XXXX**

**14. ABSTRACT**

Traditional author detection is conducted on powerful computers using documents such as books and articles. With the explosion of mobile phone computing use, modern author detection needs to be lean enough to operate on a resource restrained mobile phone and robust enough to handle the terse and non-standard wording in text messages, Tweets, and emails. By testing natural language and machine learning techniques for size and speed, not just effectiveness, this thesis identifies feature and technique combinations appropriate for author detection on a mobile phone. Specifically this thesis will examine effectiveness versus storage size for word grams of size 1, 2, and 5 as well as Gappy Bigrams and Orthogonal Sparse Bigrams. To deal with the robust nature of Tweets and text message, the Google Web1T corpus will be tested for size versus effectiveness in combination with the word grams. Once appropriate feature and technique combinations are found, those combinations will be tested on actual Android mobile phones to gauge how effective the chosen techniques are on a real mobile phone.

**15. SUBJECT TERMS**

| 16. SECURITY CLASSIFICATION OF: a. REPORT | b. ABSTRACT | c. THIS PAGE | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON / 19b. TELEPHONE NUMBER *(include area code)* |
|---|---|---|---|---|---|
| Unclassified | Unclassified | Unclassified | UU | 357 | |

NSN 7540-01-280-5500

Standard Form 298 (Rev. 8–98)
Prescribed by ANSI Std. Z39.18

THIS PAGE INTENTIONALLY LEFT BLANK

**Approved for public release; distribution is unlimited**

**AUTHOR DETECTION ON A MOBILE PHONE**

Jody Grady
Commander
Bachelor of Engineering in Aerospace Engineering

Submitted in partial fulfillment of the
requirements for the degree of

**MASTER OF SCIENCE IN COMPUTER SCIENCE**

from the

**NAVAL POSTGRADUATE SCHOOL**
**March 25, 2011**

Author: Jody Grady

Approved by: Dr. Rob Beverly
Thesis Advisor

Dr. Craig Martell
Second Reader

Dr. Peter Denning
Chair, Department of Computer Science

THIS PAGE INTENTIONALLY LEFT BLANK

# ABSTRACT

Traditional author detection is conducted on powerful computers using documents such as books and articles. With the explosion of mobile phone computing use, modern author detection needs to be lean enough to operate on a resource restrained mobile phone and robust enough to handle the terse and non-standard wording in text messages, Tweets, and emails. By testing natural language and machine learning techniques for size and speed, not just effectiveness, this thesis identifies feature and technique combinations appropriate for author detection on a mobile phone. Specifically this thesis will examine effectiveness versus storage size for word grams of size 1, 2, and 5 as well as Gappy Bigrams and Orthogonal Sparse Bigrams. To deal with the robust nature of Tweets and text message, the Google Web1T corpus will be tested for size versus effectiveness in combination with the word grams. Once appropriate feature and technique combinations are found, those combinations will be tested on actual Android mobile phones to gauge how effective the chosen techniques are on a real mobile phone.

THIS PAGE INTENTIONALLY LEFT BLANK

# Table of Contents

THIS PAGE INTENTIONALLY LEFT BLANK

# List of Figures

THIS PAGE INTENTIONALLY LEFT BLANK

# List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

# Acknowledgements

This thesis would not have been possible without the guidance and instruction of Dr. Rob Beverly and Dr. Craig Martell. Also, Dylan Freedman, who interned at NPS for the summer of 2010, did tremendous work creating the minimum perfect hash files, signature files, and scripts to create even more hash files for this thesis. His ability to grasp and implement complex hashes over a huge corpus of Google Web1T words was invaluable. Of course, the patience show by my wife Kerri-Leigh and sons, Rowan and Aiden, was a major source of support for me while creating this thesis.

THIS PAGE INTENTIONALLY LEFT BLANK

# CHAPTER 1:
# Introduction

THIS PAGE INTENTIONALLY LEFT BLANK

# CHAPTER 2:
# Prior and Related Work

## 2.1 Introduction

Author detection is the process of analyzing documents to determine if that document was created by a pre-determined set of authors.Detecting authors on mobile devices requires selection of a feature set that distinguishes authors, selection of techniques that effectively uses these features to identify authors, and selection of efficient machine learning

## 2.2 Author Detection

"Automated authorship attribution is the problem of identifying the author of an anonymous text, or text whose authorship is in doubt" [1]. For this thesis, author detection and authorship attribution are synonymous. The explosive growth of communications and document storage on the Internet provides a vast amount of data to draw on for author detection. Books, articles, blogs, tweets, and emails are posted for public viewing in an electronic format every day. Some of these postings have verifiable authors. Many Internet authors use nom de plumes or are posted anonymously. Matching verified authors to anonymous Internet authors or mobile phone texters has numerous practical applications. The increased speed and storage capacity of computing devices allow analysis of these corpora for author detection. The methods of author detection fall within the science of machine learning.

## 2.3 Machine Learning

"Machine learning is programming computers to optimize a performance criterion using example data or past experience" [2]. Machine learning has been used famously to determine the authors of the Federalist papers, allow computers to "read" human handwriting, and to mine sales data for profitable trends. Two broad categories of machine learning are supervised learning and unsupervised learning. Supervised learning is "learning with a teacher." The teacher can show the learner what to do based on examples or experience. Unsupervised learning is "learning with a critic" [2]. This thesis relies exclusively on supervised learning. Mobile device limitations demand author identification models be constructed on a platform more powerful than a mobile device. That model is then put on a device for ongoing author identification. The models require previous "teaching" instead of predictive "criticizing".

Machine learning can be used for many tasks. Often, machine learning is used to assign a given data set to a specific class or predict an outcome value over a continuous range of values. This thesis uses machine learning to assign a given data set, a document, to a given class, an author. Classification machine learning is comprised of a set of classes, a classifier, a feature set, and data. In supervised learning, the machine learner uses a data input comprised of features trained to (or owned by) by a specific class. Based on creatively counting these features, the machine learner creates a model for each class based on the the behavior of the classifier. Finally, test data, consisting of sets of features, are processed by the classifier based on the previously built models. The classifier provides an output of the most likely class that fits the given features.

Machine learning is central to this thesis. Modeling corpora of emails and tweets from numerous authors on traditional workstation or server computers, and, then, testing prediction capability on mobile devices requires not just accurate machine learning, but efficient machine learning. The efficiency is needed due to the limits of even the most advanced mobile devices. Hardware specifications are not the only limiting factor in machine learning. There are competing strengths and weaknesses in the techniques chosen, as well.

### 2.3.1 Machine Learning Techniques

The techniques in this thesis are all supervised machine learning techniques. Specifically, the two supervised techniques used are Naive Bayes and Support Vector Machine (SVM). Naive Bayes was chosen because it is computationally lightweight compared to many other methods. Support Vector Machine was chosen because data for SVM can be stored in ”sparse format”. Sparse means that every feature does not have to be represented in the stored data for a model or test case. Features with a zero count can simply be excluded. SVM has been successful in many other authorship attribution experiments [3].

#### Naive Bayes

Naive Bayes is a supervised learning method that uses Bayes Rule of probability chaining over a set of features (words in a document) to arrive at an overall probability that a specific a set of features (words in a document) belongs to a particular class (specific author). Naive Bayes uses a strong independence assumption among the various features. This means that the classifier assumes that the probability of one feature appearing in a data set is completely independent of another feature showing up in the data set. While this assumed independence of features is unlikely to be actually true, the independence keeps the calculation of probabilities simple – meaning we do not have to store the full joint set of related words and their word relation probabilities. In the case of documents and authors, Naive Bayes represents a bag of words model of a document where word order can be lost and only frequency or occurrence of words or word combination is captured. Lost word order occurs, for instance, if a unigram token is taken from a document and paired with the counts of that word in the document. In a bigram token model, word order is not lost within the bigram, but is lost between different bigram tokens. Based on a set of words, $t$, of size $n$, the probability that a document, $d$, belongs to a given class, $c$, is given by:

$$P(c|d) \propto P(c) \prod_{i<k<n_d} P(t_k|c) \tag{2.1}$$

To specifically apply the above equation to author detection, the classifier returns the class with the highest probability after executing the above formula. This turns the above equation into a maximum a posteriori (MAP) class $c_{map}$:

$$c_{map} = \arg\max_c \hat{P}(c|d) = \arg\max_c [\hat{P}(c) \prod_{i<k<n_d} \hat{P}(t_k|c)] \tag{2.2}$$

Since underflow is an issue when numerous float values are multiplied together over a set of features, the practical application of the above formula is:

$$c_{map} = \arg\max_c \hat{P}(c|d) = \arg\max_c [log\ \hat{P}(c) + \sum_{i<k<n_d} log\ \hat{P}(t_k|c)] \tag{2.3}$$

This Equation 2.3 is functionally equivalent to Equation 2.2.

Because the probability of each feature is multiplied by the probability of every other feature, a zero probability for any feature will make the overall probability zero. To handle this issue, a technique called smoothing is used. Beyond the arithmetic issue of multiplying by a zero, smoothing accounts for words the classifier has not seen due to the fact that we have incomplete data to train on. The simplest form of smoothing is Laplace Smoothing (Plus One Smoothing). In this method, each feature in the feature set is initialized with a count of 1 instead of zero. The denominator in the probability equation is increased by $1 * numberoffeatures$ to account for all the added ones. This method, attractive in its simplicity, often produces undesirable results. For this thesis, the counts from words in the Google Web1T corpus are used to smooth word counts in Naive Bayes. For example, the word ”dog” appears 3,450,297 time in the Web1T corpus, so the count for ”dog” is initialized to 3,450,297. The denominator for a Google Web1T smooth Naive Bayes instance is 1,024,908,267 based on total count weight of all tokens in the corpus.The specific details of the Google Web1T corpus are covered in a later section of this chapter.

**Support Vector Machine**

A Support Vector Machine (SVM) is a supervised machine learning method that finds a separating line or shape through a set of data based on a selected feature set. This is based on finding a boundary between two types of data in a dataset, then computing the largest boundary between closest data points and the boundary. In cases where a clear boundary between two data sets is not possible, a ”slack variable” provides an allowance of data points to be on the wrong side of the boundary. To create the boundary, SVM ”maps the input vectors into some high dimensional feature space, Z, through some non-linear mapping chosen a priori” [4]

For the two situations that a SVM can encounter: data can be separated without error and data cannot be separated without error, the same equation can be used. In the first situation, where data can be separated without error, the SVM optimizes the SVM base equation with $C = 0$. For the second situation, where the training data cannot be strictly separated, $C > 0$:

$$\min_{w,\alpha} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i}^{n}\xi \tag{2.4}$$

where $\xi$ is known as the slack variable, C is the error penalty, and the entire term $C\sum_i^n \xi$ is the soft margin. This is a quadratic programming problem to find $\xi$ and C, often accomplished by a logarithmic grid search ( $C = 2^{-5}, 2^{-3}, 2^{-1}, 2^{1}, 2^{3}, 2^{5}$ and $\xi = 2^{-15}, 2^{-10}, 2^{-5}, 2^{0}, 2^{5}$ ) with the best accuracy or F-Score determining where to continue refining the grid.

**Historical Roots of Support Vector Machines** SVM historical roots lie in the R.A. Fischer's pattern recognition work using a Variance/Covariance matrix [5]. Fisher's pattern recognition used the mean matrix (also know as the centroid of a matrix) and variance-covariance matrix (also known as the dispersion of a matrix) of two normal distributions, and found the optimal Bayesian solution was a non-linear function. Fisher simplified his non-linear function to a linear function for situations where the dispersion of both normal distributions are equal. He even found that his simplified linear equation worked satisfactorily when the distributions needing patterns recognized were not strictly normal.

From this basis, Fisher created a precedent of pattern recognition based on linear discriminating surfaces within a multi-dimensional space. Fisher's work was furthered by perceptron work in the 1960's. This work created multiple linear discriminating surfaces to find a matching pattern. However, there was no method to optimize the separation between data using perceptrons. From the need to optimize the separation, feedback mechanisms were developed to refine the perceptron weights. By further developing the idea of feeding back to the perceptron weights, SVMs were created.

### Optimal Hyperplane in Feature Space

The core of SVM is finding an optimal hyperplane in the higher dimension space mapped from the original feature space. That hyperplane is defined as:

$$\mathbf{w_0} \cdot \mathbf{z} + b_0 = 0 \tag{2.5}$$

where $w_0$ are weights, $z$ is the space, and $b_0$ is still a mystery to me. To that end, $\mathbf{w_0}$ "can be written as some linear combination of support vectors." This uses the following equation:

$$\mathbf{w_0} = \sum_{support\ vectors} \alpha_i \mathbf{z_i} \tag{2.6}$$

and the decision function using those weights is given by

$$I(z) = sign\left(\sum_{support\ vectors} \alpha_i \mathbf{z_i} \cdot \mathbf{z} + b_0\right) \tag{2.7}$$

meaning that $I(z) < 0$ for one class and $I(z) > 0$ for the other class.

For distance $\rho$ between projections defined by the support vectors, $\rho$ is defined as:

$$\rho(\mathbf{w}, b) = \min_{x:y=1} \frac{\mathbf{x} \cdot \mathbf{w}}{|\mathbf{w}|} - \max_{x:y=-1} \frac{\mathbf{x} \cdot \mathbf{w}}{|\mathbf{w}|} \tag{2.8}$$

given that 2.5 it follows that the weights needed to create the optimal hyperplane are given by

$$\rho(\mathbf{w_0}, b_0) = \frac{2}{|\mathbf{w_0}|} \tag{2.9}$$

The best solution maximizes the distance $\rho$. To maximize $\rho$, you must minimize the magnitude of $\mathbf{w_0}$. Find that minimum $\mathbf{w_0}$ is a quadratic programming issue.[4]

**Procedure** ”Divide the training data into a number of portions with a reasonable small number of training vectors in each portion. Start out by solving the quadratic programming problem determined by the first portion of training data. For this problem there are two possible outcomes: either this portion of the data cannot be separated by a hyperplane (in which case the full set of data as well cannot be separated), or the optimal hyperplane for separating the first portion of the training data is found.” If this first set is found to be linearly separable, then all the non-support vector values are discarded, a new batch of values are put into this set (these values do not meet the constraint of $y_i(\mathbf{w} \cdot \mathbf{x_i} + b) \geq 1, i = 1, ..., l$ )

**Soft Margins** In cases where the data is not linearly separable, the goal becomes to minimize the number of errors (the number of values on the wrong side of the hyperplane). Now a new variable $\xi \geq 0, i = 1, ..., l$ is introduced along with the function $\Phi(\xi) = \sum_{i=1}^{l} \xi_i^{\sigma}$ . The constraints are that the value $\xi_i$ does not push values in the non-negative quadrant out of the negative quadrant ( $y_i(\mathbf{w} \cdot \mathbf{x_i} + b) \geq 1 - \xi_i, i = 1, ..., l$. Also, $\xi_i$ is zero or a positive number ( $\xi_i \geq 0$). $\xi$ here represents ”the sum of deviations of training errors” The central equation for minimizing the number of errors is:

$$\frac{1}{2}\mathbf{w^2} + CF(\sum_{i=1}^{l} \xi_i^{\sigma}) \tag{2.10}$$

In cases for $\xi_i^{\sigma}$ where $\sigma = 1$, we are dealing with the soft margin hyperplane. Cases where $\sigma < 1$, there may not be a unique solution. For values of $\sigma > 1$, there are also unique solutions, but $\sigma = 1$ is the smallest value and that allows the term $CF(\sum_{i=1}^{l} \xi_i^{\sigma})$ from (2.10) to not overwhelm the $\frac{1}{2}\mathbf{w}^2$.[4]

**Multi-Class SVM** SVM is an inherently binary classifier. However, SVM can process multi-class data sets using SVM. There are two approaches to applying a binary classifier to a multi-class data set: one-versus-all and one-versus-one. In one-versus-all, each class in the training set is singled out against the conglomerated remaining classes in the training set. Whichever class achieves the best separation is labeled as the correct class for that data. In one-versus-one, the data classes in the training set are paired against each other and the best comparison among pairs is labeled as the correct class for that data.

It is important to define what is meant by ”best” in the classification process. Best is defined as the class that nets the most positive results from individual data instances in the training set.

Settling ties, should they occur is implementation dependent, sometimes as simple as making a random choice among the tied classes.[6].

### 2.3.2 Machine Learning Tools

There are many machine learning toolkits available. These tools come in both open source and proprietary forms. Tools are chosen based on techniques used, so, for this thesis, libSVM and libLinear were examined as SVM tools. Naive Bayes was constructed from scratch for customization with Google Web1T.

**LibSVM**

LibSVM attempts to optimize the basic SVM equation:

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\mathbf{w}^{\mathbf{t}}\mathbf{w} + C\sum_{i=1}^{l}\xi_i \tag{2.11}$$

$$\text{subject to } y_i(\mathbf{w}^{\mathbf{t}}\phi(\mathbf{x}_i) + b) > 1 - \xi_i \tag{2.12}$$

$$\text{and } \xi_i > 0 \tag{2.13}$$

For all kernels used in SVM a variable that must be solved for prior to optimization, the penalty term, $C$. Other kernels have additional variables that must be solved for prior to optimization, such as $\gamma$ in the RBF kernel. While there are sophisticated methods to find $C$ and other required variables, LibSVM takes a simple, straightforward approach: grid search. The grid for this search is a log grid search. As the local minimum is found on each pass of the grid search, libSVM reduces the grid size to home in on the minimum $C$ value.

To make libSVM more efficient and more likely to converge on a solution, data in the training set should be scaled to either span 0 to +1 or -1 to +1. While test data may show up outside the original training data range, libSVM will extend the normalized range to accommodate. For example, if the range of the training data was -100 to +100, libSVM would scale that range to -1 to +1 by dividing by 100. If there was test datum with a value of -110, then libSVM would scale that datum to -1.1. While it is stated here that libSVM scales data, that function is not automatic within libSVM itself, but is rather part of the libSVM.

LibSVM was originally constructed in C and employed with python tools to support. LibSVM is not available in a wide array of languages, including Java. A Java version of libSVM makes libSVM functional on many of the mobile operating systems available today, including Android. For this reason, libSVM was originally chosen as the SVM tool for this thesis.

**LibLinear**

While libSVM has numerous kernels to improve results, the inclusion of code to accommodate these kernels slows libSVM down. To increase processing speed for libSVM for linear kernels, libLinear was created. LibLinear is heavily modeled on libSVM but without non-linear kernel support. The kernels, represented within the $\phi$ function in SVM equations is not dealt with at all in libLinear, thus cutting down on checks and processing time. A linear kernel has been found to give as good or nearly as good a result as other kernels such as RBF, parabolic, and radial for text classification, especially when the corpus being used is large. The reduction in code can produce results 100-200 times faster that using LibSVM.

LibLinear has also been studied for large data sets that produces models which cannot be fit into memory. the application of ”chunked” data on a mobile platform with very limited RAM, but significant storage (due to microSD cards) makes libLinear even more attractive for mobile device use.

## 2.4 Features

### 2.4.1 Feature Types

Feature types for natural language processing can be as simple a keeping counts of individual characters within a document to complex tracking of word combinations. There are three feature types used in this thesis, N-Grams, Gappy Bigrams, and Orthogonal Sparse Bigrams. These feature types vary in complexity and effectiveness for author detection.

**N-Grams**

N-grams are word groups or character groups of size N within a document. These word groups can include sentence boundaries, often denoted as $<$ S $>$ for sentence start and $<$ /S $>$ for sentence end. For instance, in the phrase the ”the quick brown fox” the set of 2-grams (bigrams) are shown in Table 2.1:

To further illustrate, the 3-grams (N=3 N-grams) of the phrase ”the quick brown fox” are show in Table 2.2:

Table 2.1: N-grams (N=2) of "the quick brown fox" with sentence boundaries

| |
|---|
| $<$ S $>$ the |
| the quick |
| quick brown |
| brown fox |
| fox $<$ /S $>$ |

Table 2.2: N-grams (N=3) of "the quick brown fox" with sentence boundaries

| |
|---|
| $<$ S $>$ the quick |
| the quick brown |
| quick brown fox |
| brown fox $<$ /S $>$ |

The larger the N-Gram, the lower the probability of finding that N-Gram in a document. A specific 5-Gram may be very rarely repeated, even by the same author. That makes a 5-gram distinctive, but unreliable for author detection. A 1-Gram like "the", "of", "a", etc occurs frequently across almost all authors, but is not discriminating. Finding discriminating words groupings without the unreliable low probability of large-N N-Grams drove the creation of a modified N-Gram grouping called a Gappy Bigram.

**Gappy Bigrams**

Gappy Bigram definitions vary between the sources cited in this thesis. For the purposes of this thesis, a Gappy Bigram will be composed of two tokens (words) found within a distance of words. A Gappy Bigram of distance 0 reduces to an identical set to 2-Grams (also know as bigrams). Just like N-Grams, Gappy Bigrams can extend beyond a sentence boundary, include punctuation, etc. However, for larger distances, the distinction between Gappy Bigrams and regular bigrams is clear. For instance, in the phrase "the quick brown fox" and a Gappy Bigram distance of 2, the Gappy Bigrams are given in Table 2.3:

To further illustrate, Gappy Bigrams of distance 1 are given in Table 2.4:

The Gappy Bigram is able to preserve distinctive word groups for an author without the extremely low probability of occurrence. However, an author may distinctively used a two word group at exactly an interval of 3 words or 2 words or 1 word. That distinctiveness could be a key attribute for that grouping and is lost in Gappy Bigrams. To capture that distinctiveness, Orthogonal Sparse Bigrams are employed.

Table 2.3: Gappy Bigrams (of distance 2) of ”the quick brown fox” with sentence boundaries

| |
|---|
| $<$ S $>$ the |
| $<$ S $>$ quick |
| $<$ S $>$ brown |
| the quick |
| the brown |
| the fox |
| quick brown |
| quick fox |
| quick $<$ /S $>$ brown fox |
| brown $<$ /S $>$ |
| fox $<$ /S $>$ |

Table 2.4: Gappy Bigrams (of distance 1) of ”the quick brown fox” with sentence boundaries

| |
|---|
| $<$ S $>$ the |
| $<$ S $>$ quick |
| the quick |
| the brown |
| quick brown |
| quick fox |
| brown fox |
| brown $<$ /S $>$ |
| fox $<$ /S $>$ |

**Orthogonal Sparse Bigrams**

Orthogonal Sparse Bigrams (OSB) are similar to Gappy Bigrams in how there are constructed except that the distance between words in the OSB is included in the OSB. Just like N-Grams, Gappy Bigrams can extend beyond a sentence boundary, include punctuation, etc. For instance, in the phrase ”the quick brown fox” and a OSB distance of 2, the OSBs are given in Table 2.5:

To further illustrate, OSBs of distance 1 are given in Table 2.6:

It is important to note that in the cited references, the distance for OSBs is place between token 1 and token 2 instead of after token 1 and token 2 as shown in Tables 2.5 and 2.6. The distance is placed after the tokens in this thesis for more convenient parsing within reference files. Also, for OSBs, there is an issue of how to count OSBs. The two approaches are to strictly use on the distance that a token pair is found. In ”the quick brown fox”, the OSB of distance 2 of ”quick

Table 2.5: Orthogonal Sparse Bigrams (of distance 2) of "the quick brown fox" with sentence boundaries

| |
|---|
| < S > the 0 |
| < S > quick 1 |
| < S > brown 2 |
| the quick 0 |
| the brown 1 |
| the fox 2 |
| quick brown 0 |
| quick fox 1 |
| quick < /S > 2 |
| brown fox0 |
| brown < /S > 1 |
| fox < /S > 0 |

Table 2.6: Orthogonal Sparse Bigrams (of distance 1) of "the quick brown fox" with sentence boundaries

| |
|---|
| < S > the 0 |
| < S > quick 1 |
| the quick 0 |
| the brown 1 |
| quick brown 0 |
| quick fox 1 |
| brown fox 0 |
| brown < /S > 1 |
| fox < /S > 0 |

brown" has one instance, with a distance of 0. For the other approach, using lesser-included distances for OSB of distance 2 , "quick brown" has three instances: quick brown 0, quick brown 1, and quick brown 2 because quick brown is a lesser included OSB of distance 2.

If file or database of OSBs is constructed, then a file or database of Gappy Bigrams also exists by default. The count of maximum distance OSBs equals the count of Gappy Bigrams, assuming the lesser included version of OSBs is used. This can be useful for conserving space in a system when both OSBs and Gappy Bigrams are needed.

### 2.4.2 Feature References

Once a scheme is determined for managing features, the features required must be selected. Feature selection is the process of deciding which features to include during classification. A set of features can be built from the training set, such as selecting N most used words in a training set. Features can be further refined by using outside references. For instance, a feature set could be built as the N most used words in a training set and filtered for stop words. In this case, stop words could be defined by other researchers work or some standard stop word set. Another option is to build all features from a reference set. This thesis made heavy use of the Google Web1T Corpus to act as a feature filter and a feature reference.

**Google Web1T Corpus**

The Google Web1T Corpus is a massive corpus of English language N-grams ranging from N=1 to N=5. The corpus was created from a snapshot of Google's search databases that took place during January 2006. The corpus consists of text files with the N-grams accompanied by a count of those N-grams. Each set of N-grams is stored in its own folder. The N-Grams are organized alphabetically by the first word in the N-Gram. For instance, ”a cat” comes before ”a dog” in the 2-Grams of the corpus.

The unique folder within the corpus is the 1-Gram folder. There are two files within the 1-Gram folder. One file is organized alphabetically like the rest of the corpus, but the other file is organize by count. The largest count comes first. This folder serves as both a 1-Gram source and an authoritative reference of all types within the Google Web1T corpus.

Punctuation is included in the corpus. Sentence boundaries are indicated by <S> and < \S >. To qualify for corpus inclusion, a 1-Gram needed to appear in the Google search databases at least 200 times. Additionally, to appear in a 2-Gram or greater, a gram had to appear in the database at lest 40 times. For 2-Grams and greater that appeared 40 times or more, but one of the words in the gram did not individually appear at least 200 times, the tag <UNK> is used to replace that word. The characters used in the corpus are UTF 8. Tokenization was ”similar” to Penn Tree Bank except that hyphenated words were separated.[7] From working with the corpus, it becomes apparent that contraction within the corpus does not exactly match Penn Tree Bank. No ”'t” contractions were kept intact during tokenization. The authors were contacted regarding this tokenization issue, to determine if this was intentional, but no reply has been received.

The Google Web1T is massive. This size makes Web1T both powerful to employ and cumbersome to use. The statistics for this corpus are listed in Table 2.7.

Table 2.7: Token and Type Counts in Google Web1T Corpus

| | |
|---|---|
| Number of tokens: | 1,024,908,267,229 |
| Number of sentences: | 95,119,665,584 |
| Number of unigrams: | 13,588,391 |
| Number of bigrams: | 314,843,401 |
| Number of trigrams: | 977,069,902 |
| Number of fourgrams: | 1,313,818,354 |
| Number of fivegrams: | 1,176,470,663 |

### 2.4.3 Feature Compression Techniques

Due to the large size of the corpora and feature reference used in this thesis, an efficient way to represent words and N-grams was needed. After surveying general literature on representing large data sets, the search for this thesis was narrowed. Two methods of efficiently representing large sets were investigated: bloom filters and minimal perfect hash functions. Minimal perfect hash functions were ultimately chosen as the tool for representing data in this thesis.

#### Bloom Filters

Representing a large dataset in a small memory space requires trading off between probability of a false positive, probability of a false negative, processing time, and size of representation. Bloom filters allow efficient storage of a list of values with zero probability of false negatives and a minimum probability of false positives. A Bloom filter consists of an array of $m$ bits and $k$ hash functions. Each hash function has an output range of 0 to $m-1$. Each hash function must provide an equal probability distribution for each value 0 to $m-1$. At the beginning of the construction of the Bloom filter, all $m$ bits are set to 0. Each value to be a member of the Bloom filter is processed by each hash function. The output of the hash function corresponds to the array position of one of the $m$ bits, which is then set to 1. If an output bit is already set to 1, that bit remains a 1. After all Bloom filter member values have been processed by the hash functions, the array of bits should be a mix of 0's and 1's.

To determine if a value belongs to the Bloom filter, that value is run through all $k$ hash functions. If each array position output by the $k$ hash function contains a bit set to 1, then the value probably belongs to the Bloom filter. If any of the m bits is a 0, that value does not belong.

There are variations on the Bloom filter that can use parallel architectures to advantage. For example, if the array of $m$ bits is a multiple of $k$, then each hash function can have a range of 0 to $\frac{m}{k}$. Then each hash function can be run in parallel instead of in series. This scheme has no effect on the probability of a false positive, but can be appreciably faster to process in parallel processing platforms.

The work in a Bloom filter comes from determining the minimum values required for $k$ and $m$ to represent the expected set of values for a required false positive rate. The trade offs are, the larger the number of bits, the lower the probability of a false positive, but the larger the storage of the Bloom filter becomes. Likewise, an increased number of hash functions provides a lower probability of false positives, but larger numbers of hash functions increases the computational cost of the Bloom filter. Given a required maximum false positive probability, $p$, and a maximum number of items, $n$, the minimum number of bits, $m$, is given by:

$$m = \frac{n \ln p}{(\ln 2)^2}. \tag{2.14}$$

Once the number of bits, $m$, is determined, the minimum number of hash functions, $k$, must be found. The required minimum of hash functions is given by:

$$k = \ln \frac{m}{n}. \tag{2.15}$$

Bloom filters are flexible and compressible. They are flexible because the number of bits, $m$, can be changed on the fly based on a changing number of items, $n$. Various compression techniques can be used to compress the bits, $m$, in the filters for transmission between computers. The filters can be processed in serial or parallel based on hardware architecture. Unfortunately, while flexible, Bloom filters are not as compact as their closely related cousin, the minimal perfect hash function.

**Minimal Perfect Hash Functions**

A minimal perfect hash fuctnion is the culmination of three concepts: a hash, a perfect hash, and a minimal hash. A hash function is a function that maps values from a set, $U$, with a number of values, $k$, to a range of values, $m$ [8]. Hashes are normally associated with mapping a large universe to a small universe, but hashes can map between spaces of equal size. Hashes are often used in computer science for cryptography, efficiently mapping values, and myriad other tasks.

A hash function is a perfect hash function if there are no hashing collisions. A collision occurs when different values from U result in the same output value. More formally, in perfect hashes, there are $m$ distinct values resulting from applying the hash function to all $k$ values in $U$ such that $k = m$. In short there must be a 1-1 mapping between each value in $U$ to each resulting value in the range, $m$ – no collisions to be handled (load factor $\alpha = 1$. A perfect hash function is called a k-perfect hash function if the ratio of possible values in the mapped space is not larger than k times the original space. This means the range, $m$, must be $k$ time larger than U to ensure there are no collisions.

A perfect hash function is called a minimal perfect hash function if there are no "blank" spaces in the hash table – meaning that no space is wasted in storing the hash. This is the same as a k-perfect hash function where k=1. Less formally, the size of the range, $m$ is equal to $n$, the size of the universe, $U$.

The time required to compute a value in $m$ from a value in $U$ is known as evaluation time. The required to construct the minimal perfect hash function is known as construction time. Along with representation space, evaluation time and construction time are the three performance parameters used to judge the efficiency of a minimal perfect hash function.

Minimal perfect hash functions (MPHF) are comprised of a set of hashing functions and a lookup data structure. The set of values (the universe, $U$)to be hashed must be known in advance. Those values are mapped, one-to-one to a unique range of numbers. At the end of the mapping, there is exactly one unique numerical hash for every provided input. The required number of bits for the hash is the minimum number of bits possible to uniquely identify all the items. The theoretical lower bound is 1.44n bits, where $n$ is the number of elements in $U$.

A lower bound of 1.44n bits is the advantage of the MPHF, the data structure is extremely compact once created. The disadvantage is that any value submitted to the MPHF will result in a hash value. This requires a second discriminating function to determine member in the correct value set, such as a second, traditional, hash. This second hash undermines the compact size of the MPHF. However, combining a MPHF with a single traditional hash provides an extremely small probability of a false positive during a membership check and a fast lookup time.

In general, there are three stages of creating a minimal perfect hash function or any k-perfect hash function. These three stages are mapping, ordering, and searching. The mapping stage maps the set of keys in universe, $U$, to some other values. For example mapping a set of strings to an integer value or creating a set of vertices in a graph could serve as the mapping step. Ordering involves finding the buckets, vertices, etc that have been mapped with the most keys. These highly mapped entries become levels or child graphs in a further refined hashing scheme to develop into the final data structure. The final step, searching, involves assigning keys to positions within the mapping. The mapping is often multilevel allowing duplication from hashing to be ”backed off” and retried to continue building the hash.

There are many MPHF implementations available in the open source world. The implementation claiming to be the closest to the theoretical minimum for representation space is called the Compress, Hash, and Displace (CHD) algorithm[9]. CHD maps keys into buckets. Each bucket is assigned its own hash function, $\phi$ to create an index into the final data structures. The buckets are ordered by magnitude (number of values in the bucket) for placement into the data structure. The theoretical lower bound of storage for a minimal perfect hash is 1.44n bits [8]. CHD's lower bound of storage is 2.07n to 3.56n bits depending on generation time allowed for the data structure.

## 2.5 Evaluation Criteria

Results from classifying data are computed from four basic categories of results: true positives, $(tp)$, true negatives $(tn)$, false positives $(fp)$, and false negatives $(fn)$. These four basic results are combined into accuracy, precision recall and F-Score for this thesis. While there are other evaluation criteria, these chosen criteria are clear enough and sufficient for measuring results.

### 2.5.1 Accuracy

Accuracy is a widely used and intuitive performance measure for classification. Accuracy, however, is flawed. Accuracy poorly represents the effectiveness of a classifier when the number of true negatives is large compared to the number of true positives. Missing all the true positives, but calling everything a negative, true or otherwise, yields a high accuracy without actually being effective at finding correctly labeled positives. Accuracy is defined as:

$$accuracy = \frac{tp + tn}{tp + fp + tn + fn} \tag{2.16}$$

[10]

### 2.5.2 Precision and Recall

Due to the weakness of accuracy as an evaluation criteria, precision and recall (also known as sensitivity) is used. Precision measures how often a document that belongs to the class being sought is actually labeled as that class. In other words, for all the actual documents written by the target author, how often are those documents labeled by the classifier as being written by the author. For all the documents said to be true by the classifier, what percentage are actually true.

$$precision = \frac{tp}{tp + fp} \tag{2.17}$$

Recall determines how well the classifier picks out true documents. In other words, for all the true documents in the set, how often does the classifier detect those true documents? Recall is given by:

$$recall = \frac{tp}{tp + fn} \tag{2.18}$$

[10]

### 2.5.3 F-Score

F-Score is the harmonic mean of precision and recall. It is a superior indicator to accuracy in evaluating a classifier. The definition of F-Score used in this thesis is:

$$F - Score = \frac{2}{\frac{1}{p} + \frac{1}{r}} \tag{2.19}$$

This definition is a variant of the standard definition of:

$$F - Score = \frac{(\beta^2 + 1) * 2pr}{\beta^2 * (p + r)} \tag{2.20}$$

The full definition of F-Score involves an additional term, $\beta$, which is a weighting value. A $\beta$ value greater than one favors precision and a $\beta$ value less than one favors recall. This thesis values precision and recall equally. This makes $\beta = 1$, thus the simpler equation for F-Score is used:

$$\frac{2pr}{p + r} = \frac{2}{\frac{1}{p} + \frac{1}{r}} \tag{2.21}$$

F-Score will be the primary evaluation criteria for this thesis.[10]

## 2.6 Mobile Device Platforms

There are numerous mobile device platforms ranging from the near ubiquitous mobile phones to tablets to personal digital assistants. Even within the category of mobile phones, there is a wide ranging array of capability and popularity. For newer mobile phones, capabilities often include access to storage, a network, phone services, GPS, and multimedia. Storage can be both onboard phone storage or removable storage such as a micro-SD card.

Often, there network access to more than just the mobile provider GSM or CDMA network. Modern phones often have WiFi access. GPS services provide position updates to the phone. Multimedia capability varies dependent on display size, resolution, battery consumption, processing speed, memory, and network availability. Mobile phones have not yet reached the level of commonality expected in desktop and laptop computing devices.

### 2.6.1 Mobile Devices by Popularity

To determine an effective development strategy for author detection on a mobile phone, it is sensible to determine what development language would support the largest number of mobile phones. By device popularity, the most dominant mobile operating systems, in order, are Symbian (Nokia phones), Research In Motion (Blackberry), iOS (Apple iPhone, iPad, iPod), and Android (Droid, Evo, Galaxy Tab). These four OS platforms constitute 88% of the mobile device market for first quarter of 2010.[11] Symbian, RIM, and Android all accept applications built on Java, or at least a variant of Java. Based on this vast market share, using Java as the development language for author detection on a mobile device has the largest potential for use.[12][13][14] Only iOS uses exclusively Objective C.[15]

### 2.6.2 Android Operating System

Based on its popularity and ease of installing test applications, Android is used as the development platform for this thesis. Android applications are not written, strictly speaking, in Java. Android applications are written in Dalvik which implements most of the syntax and structure of Java. Dalvik development is targeted at mimicking recent stable releases of the Java Development Kit (JDK). The core of the Android operating system is built on Linux, but is not built as a traditional Linux environment.[14]

Android applications consist of a combination of Activities, Services, Intents, and Content Providers. Activities are processes that users can see and interact with. Activities create the windows, tabs, and dialogs for user interaction.

Services run in the background with no user graphical user interface (GUI). Android Services are not equivalent to traditional Unix services. Unix services are, by nature, persistent process within the operating system. Android Services are just as prone to being killed by the operating system as an Activity.

Intents are messages passed around by processes and Java Virtual Machines within the Android operating System. Typical Intents are created by Content Providers for actions such as incoming calls, incoming Short Messaging Service (SMS) messages, GPS, etc. Other typical Intents are passed between Activities in an application or between Services and Activities in an application. Intents can start, stop, and pause Activities as well as just pass along data such as a String or integer. Applications use Activities, Services, and Intents in combination to provide functionality on an Android Mobile device.

The lifecycle of an application in Android varies from a standard PC application lifecycle. Activities and Services continue to run in Android while sufficient resources remain on the mobile device. When resources become exhausted, the Android operating system will shut down Activities and Services it deems as less important or less used. This is why Android applications often lack a ”Quit” or ”Exit” function in their menus – developers expect that the application can continue to run so long as the operating system has sufficient resources. Contents providers, on the other hand, are persistent processes driven by items such as GPS receivers, mobile networks, and WiFi networks. Content providers are accessed and listened to by applications. A Content Provider can also be built by a developer to act as a data provider for other application as an abstraction instead of an actual physical device like GPS or WiFi.[16]

## 2.7 Corpora

A major portion of validating a method of author attribution is securing a corpus of usable data. There are some tried and true corpora openly available, such as the ENRON Email Corpus, which are well know, well studied, and useful for comparison. With a focus on mobile devices, this thesis needed a more short text relevant corpus. For this need an in-house corpus of Twitter posts, known as Tweets, was used. Using these two corpora provides a standard corpus to judge effectiveness and a newer corpus to anticipate future capability in the evolving medium of mobile computing.

### 2.7.1 ENRON Email Corpus

The ENRON email corpus is a set of emails collected by the Cognitive Assistant that Learns and Organizes (CALO) Project. The original corpus contains 619,446 emails from 158 users. These emails were posted on the web by the Federal Energy regulatory Commission during the investigation of ENRON. Issues with the raw posting were corrected by several people at MIT and SRI to arrive at the form of the current corpus. The emails are organized in folders, by user. The folder organization used by the original user is kept mostly intact (Inbox, Sent Items, etc) except for some computer generated folders that were seldom used by the actual users. Each email is contained in its own text file. Each text file contains the full email header as well as any threaded conversation headers (replies and forwards).[17]

The ENRON corpus is a frequent target for natural language processing. Author detection performance for character and word N-grams, SVM, Naive Bayes and other classifiers on the ENRON corpus is well documented. For this reason, all methods used in this thesis were attempted on the ENRON email corpus as a benchmark of performance, before moving on to the more mobile-centric corpus of Twitter.

### 2.7.2 Twitter

Twitter is a short message micro-blogging services that users can access from traditional computers as well as mobile devices. Originally designed for use over Short Message Service (SMS), Tweets (vernacular for message sent on Twitter) are limited to 140 characters. Unlike other social networking sites, Twitter has no requirement for users to post their real names. Author detection on a corpus of Tweets will be challenged by the short duration of each Tweet (Tweets would constitute a document in this case) and the non-standard use of language. Also, users do not have to formulate original content for their Tweets. Just like as email forward, users can re-Tweet a Tweet they have already received.

Tweets are formatted for use with a JavaScript Object Notation (JSON) format. The JSON formatting provides numerous fields containing language, Twitter id, geocode (latitude and longitude of sender). The Twitter API contains both streaming and RESTful methods. Using the Twitter API, Tweets can be pulled from the TwitterSphere using a free, rate limited service called Garden Hose or via a fee-based, rate unlimited service called Fire Hose. The rate limit for Garden Hose is 150 messages per hour. Those messages are randomly chosen from Twitter accounts that make themselves viewable by the public. The Twitter API allows for filters to affect the stream of Tweets to avoid getting Tweets that do not meet your needs and would otherwise impact your rate limit. The length limitation and mobile nature of Tweeting, makes Twitter a reasonable model of SMS behavior for testing purposes.[18]

## 2.8 Recent Work in Author Detection, Google Web1T, and Mobile Devices

:NOTE: I just found a slew of related work that is worth studying some more, but I don't want to hold up the process of getting feedback on the bulk of this chapter 2. The below incomplete sentences are placeholders for me on what I have for related work right now. There is a patent on author detection based on a compressed hash, there are articles solely on managing and querying the Web1T, and a paper on chunking data on memory constrained systems. :ENDNOTE:

Google Web1T has been used as a smoothing reference in other machine learning studies. XXXXXX used a backoff method based on Google Web1T XXXX counts to ....... Google Web1T has also been a reference for spelling correction[] and semantic classification[]. There has even been a paper on just managing the Google Web1T corpus effectively[].

Author detection across varied information sources using a normalized compressor distance has been patented. This method creates a bitwise compression of content from web pages, emails, texts, or any electronic document and uses clustering, based on this patented distance measure, to arrive at probability of various documents being from the same author.[**?**] Author detection on mobile devices has not shown up in patent or paper searches. There are author detection papers that reference the prevalence of text messages in author attribution, but none on using the mobile platform itself to conduct processing. Despite a breathtaking pace of application development on mobile device platforms such as iPhone and Android, using mobile computing capability for traditional machine learning has appears to be a wide open question.

Recent SVM work has included....

## 2.9 Conclusion

There is a rich body of work on author attribution, SVM, Naive Bayes, and on the ENRON Email corpus. Applying traditional document and email author attribution tools to the short message environment of mobile phones is an area ripe for exploration.

THIS PAGE INTENTIONALLY LEFT BLANK

# CHAPTER 3:
# Experimental Design

This chapter document the concepts and technical approaches used in this thesis, as well as procedural concepts for understanding the experiments of this thesis.

## 3.1 Experimental Design Overview

**Thesis Goals** The central goal of this thesis's experiments is to compare size and speed of different author detection methods against the effectiveness of those same author detection methods on a resource constrained device such as a mobile phone. Size and speed are critical to this thesis. This is due to the restrictive nature of mobile phones. However, the nature of these experiments allows the results to be applied to other computing platforms with limited resources such as nano-computers, mobile sensors, or yet unimagined devices.

**Experimentation Phases** To achieve the thesis goal, experimentation will be conducted in two phases: parameter evaluation and mobile phone performance evaluation. In parameter evaluation, the effectiveness of different combinations of classification methods, features sets, group sizes, and smoothing/filtering to compare prediction performance against model size and processing requirements. During the mobile phone performance evaluation, the combinations of classification methods that are both feasible and effective are used on mobile phones to determine the overall performance and impact of running author detection on an actual mobile phone.

## 3.2 Phase One: Parameter Evaluation

This phase will evaluate numerous combinations of two classification methods, five feature sets, six grouping sizes, three grouping methods, and two corpora to determine the computing requirements and effectiveness of these combinations. Preparing for these evaluations takes several steps including determining the required combinations, organizing and compressing the feature references, preparing the training and prediction data, building the models, and, finally, running the prediction tests. The results for all prediction test will be stored in a mySQL database which will also store the resulting f-score, precision, recall, and size of model for each test.

### 3.2.1 Creating the Testing Combinations

The classification methods to be compared are Naive Bayes and Support Vector machines (SVM). Naive Bayes is fast and uses a relatively small amount of RAM and disk storage. SVMs, are slower, use greater RAM and disk storage, but often yield higher f-scores. There are numerous feature sets that can be chosen. For this thesis, 1-grams, 2-grams, 5-grams, gappy bigrams, and orthogonal Sparse bigrams will be examined. The intuition is that 1-grams are simple and use less space, but will be less effective than bigger feature sets such as gappy bigrams or 5-grams.

For this thesis, two feature reference sets will be examined, a bootstrapped bag of words and the Google Web1T corpus. Bootstrapped bag of words simply means finding all the unique types within a training set and making each type a feature in the feature set. Since the Google Web1T corpus is huge, a parameter of that feature reference which can be adjusted is the percentage of a given feature set that might be used. These experiments will permute through these numerous options to determine size, speed, precision and f-score. The end result will be an analysis of the utility of these various approaches to author detection on a mobile phone. A graphic of the parameter combinations is given in Figure 3.1.

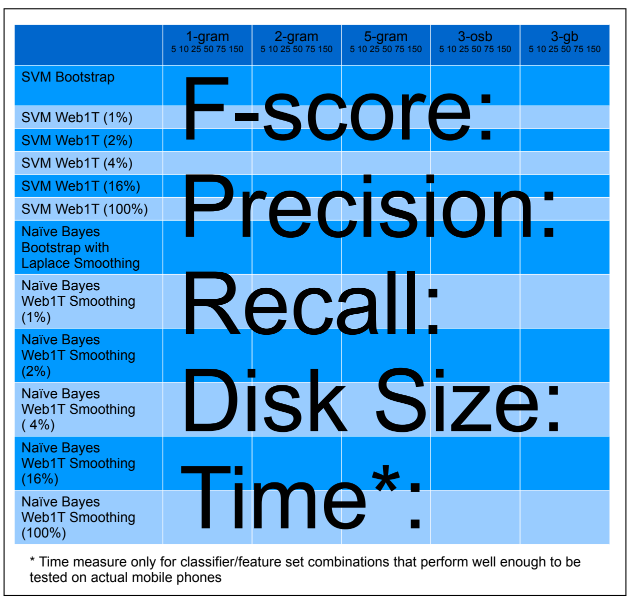



Figure 3.1: Parameter Combinations for Testing

The small numbers ”5 10 25 50 75 150” given under each column heading in Figure 3.1 indicate that all authors will be tested in groups of 5, 10, 25, 50, 75, and 150 using three different grouping strategies: small-to-large, small-and-large, and random. In small-to-large, the authors with the smallest amount of training data are grouped together. In small-and-large, small authors and large authors are paired together. In the random grouping, the authors are grouped together by a pseudo-random selection. The reasoning for these three grouping strategies is to provide insight into the effect of prolific authors versus less prolific authors. If results are similar for the same author for each group, the prolific writing may not impact the outcome of author detection with these methods. This is needed information to rule out that the test author detection methods simply select the most prolific author instead of the actual author.

### 3.2.2 Organizing and Compressing Feature References

A key element to this testing is the use of the Google Web1T corpus. The Web1T corpus contains billions of types with a token mass of just over 1 trillion. The size and breadth of the Web1T corpus makes it appealing as a source for smoothing in Naive Bayes and a tool for creating models in SVM. However, due to the huge size of the Web1T corpus, the text files comprising the corpus must be compressed and managed for use on desktop workstations, servers, and especially mobile devices. Managing the corpus requires determining what portions of the Web1T corpus will be used. Using the choice of 5-grams as an example for illustration purposes, suppose only the 5-grams portion of the Web1T corpus might be used. The 5-grams constitutes 118 text files containing up to 10 million lines of text each. Each line of the Web1T 5-gram files contains space separated words (making up the type) followed by a count, separated from the words by a tab. The lines of text are organized alphabetically by token where uppercase letters are distinct from lowercase letters. Even using only one size of gram from Web1T, a reverence of this size is slow and bulky for machine learning use. Therefore, a subset of the reference is needed.

**Sizing the Feature Reference Set** To manage the size of Web1T, a small portion of the 5-grams could be chosen – 1%, 2%, 4%, etc. To choose which part of the reference to use (largest, smallest, random) this thesis takes advantage of Zipf's Law. Zipf's law states that the highest frequency word occurs approximately twice as often as the next most frequent word. By that reasoning, a list of the types with the highest counts is needed to capture the largest use of words in a natural language corpus. To get this count ordered list, the complete set of Web1T n-grams are recreated. The recreated files list each type organized by count instead of alphabetically. If two or more types have the same count, then those types are list alphabetically. The types are still listed first as a group of space separated words followed by a tab and ended with a count.

**Three Tiered Hashing Scheme** Even once the feature set of types to be used for classification has been determined, the smaller set of text is still too slow to process and very bulky to store. To further compress the data, a three tiered hashing scheme is used. The structure of the three tiered hashing scheme is shown in Figure 3.2.

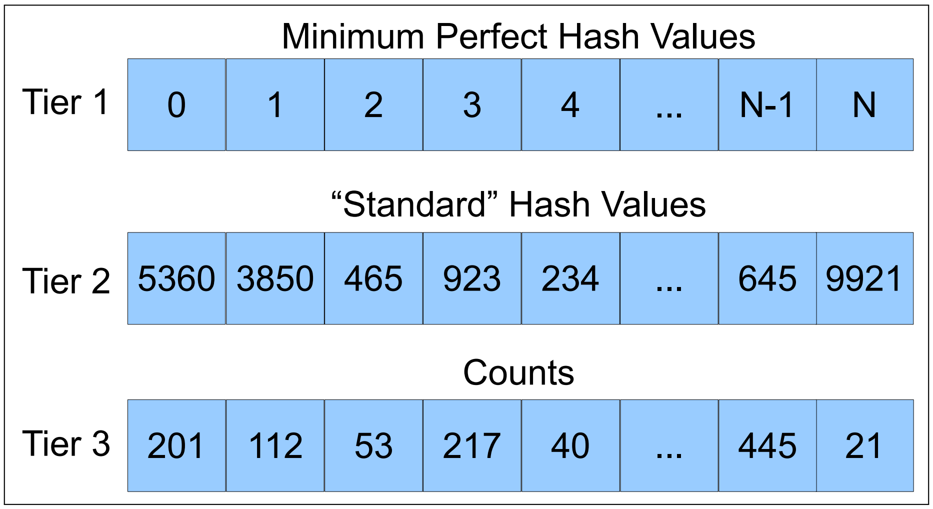



Figure 3.2: Three Tiered Hashing Scheme Structure

The first tier is comprised of minimal perfect hash (MPH) values of the selected feature set. The second tier of the scheme is comprised of a 64 bit hash of the original type. This second tier's job is to reduce the probability of a false positive in the fist tier. This issue arises because no matter what string is input to the MPH function, a valid MPH value will be produced. The second tier's traditional hash is accessed by mapping the MPH value to the index of an array that comprises the second tier. That array cell contains the 64 bit hash of the original text used to create the MPH value. This make the false positive rate for a given type $\frac{1}{2^{64}} * \frac{1}{\text{range of MPH values}}$ which is deemed an appropriate risk of collision in this hashing scheme. The third tier is simply an array of long values. The MPH value from tier 1 is used to access this array which hold the count value for a given type. An example of converting a phrase, ”the quick brown”, is shown in Figure 3.3.

These different tiers are not contained in a single data structures. The MPH data structure, tier 1, is contained in a file called ”keys.mph”. The array of hash values, tier 2, is contained in a file called ”signature”. The counts are contained in a Java object file call LongCountsArrayFile. The Naive Bayes experiments use all three tiers of this structure for smoothing values. The SVM experiments only use tier 1 and tier 2 to verify that a string encountered actually belongs to the

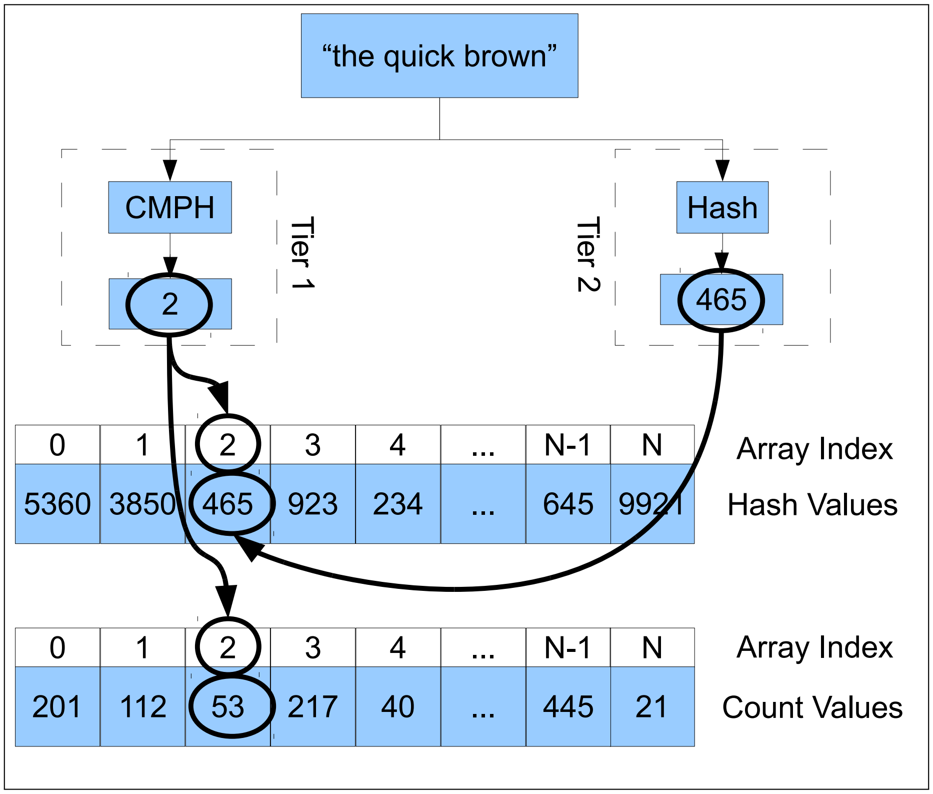



Figure 3.3: Three Tiered Hashing Scheme Example

feature set. These hefty data files comprise the bulk of storage required on the mobile device. Since these data files get loaded into RAM during the prediction process, the file sizes also impact RAM requirements. The impact on RAM and disk storage makes management of the size of keys.mph, signature, and LongCountsArrayFile an important aspect of the experiments.

**Choosing Artifacts for the Three Tiered Hashing Scheme** One impact of using MPH to reduce the size of storing types is a loss of flexibility with the text artifact selection process. Before the MPH data structure is created, the creator must determine if punctuation, capitalization, sentence boundaries, or "unknown" words will be allowed. The omission of each of these artifact types brings its own unique challenges. A binary style number scheme was adopted for each of these features where capital letters hold the 1 position, punctuation the 2 position, unknown word tags the 4 position, and sentence boundaries hold the 8 position. The complete matrix of artifacts allowed in the MPH model is included in Figure 3.4.

| MPH Label | Remove Sentence Boundary Tags | Remove Unknown Word Tags | Remove Punctuation | Remove Capital Letters |
|---|---|---|---|---|
| 0 | FALSE | FALSE | FALSE | FALSE |
| 1 | FALSE | FALSE | FALSE | TRUE |
| 2 | FALSE | FALSE | TRUE | FALSE |
| 3 | FALSE | FALSE | TRUE | TRUE |
| 4 | FALSE | TRUE | FALSE | FALSE |
| 5 | FALSE | TRUE | FALSE | TRUE |
| 6 | FALSE | TRUE | TRUE | FALSE |
| 7 | FALSE | TRUE | TRUE | TRUE |
| 8 | TRUE | FALSE | FALSE | FALSE |
| 9 | TRUE | FALSE | FALSE | TRUE |
| 10 | TRUE | FALSE | TRUE | FALSE |
| 11 | TRUE | FALSE | TRUE | TRUE |
| 12 | TRUE | TRUE | FALSE | FALSE |
| 13 | TRUE | TRUE | FALSE | TRUE |
| 14 | TRUE | TRUE | TRUE | FALSE |
| 15 | TRUE | TRUE | TRUE | TRUE |

Figure 3.4: Matrix of CMPH Models by Artifacts Included

**Omitting Punctuation** Omitting punctuation provides two options for dealing with the corpus: replace punctuation with "$<$ UNK $>$" or drop the punctuation altogether. If punctuation is dropped, then any type containing a punctuation mark in the feature reference set must be completely ignored. If the punctuation is replaced with $<$ UNK $>$, then a search within the existing count structure must be conducted for a corresponding entry for $<$ UNK $>$ and any non-punctuation words in the type. While dropping punctuation is much simpler to implement than employing "$<$ UNK $>$" tags, however, Google did count punctuation as a word in type construction, so correlation between n-gram counts in the Web1T corpus and the trained/predicted documents is slightly affected. To maintain simplicity, the simple drop approach was used in these experiments.

**Omitting Capitalization** Omitting capitalization is straightforward for construction of tier 1 and tier 2, the inputted text for the type is converted to all lower case and a check is conducted to see if that type is already in the MPH data structure. For tier 3, which contains the counts, the lower case versions of the word must have its count mass added with its corresponding uppercase types. This adds complexity to the insertion process for MPH but is easily managed. Another option would be to simply drop all types that contained capitalization, but that would remove a large count mass from the Web1T corpus. Adding counts was the method used in this thesis to deal with omitting capitalization.

**Omitting Sentence Boundaries** Sentence boundaries are denoted in the Web1T corpus as $< S >$ and $< \backslash S >$. Dropping sentence boundaries is straightforward since there is no replacement or count mass issues to deal with. Since the tools for locating sentence boundaries make use of their own machine learning processes, no sentence boundaries were used in these experiments.

**Omitting Unknown Words** In the Web1T corpus, ”unknown” words have a specific meaning. To be included in any corpus n-gram set, a word must have appeared as a 1-gram at least 200 times in the Google database. By contrast, to be 2-gram, 3-gram, 4-gram, or 5-gram, that gram had to appear at least 40 times in the Google database. This created as situation where a word would need to appear in a 2-or-higher-gram, but was not allowed into the corpus because it did not appear 200 times in the overall database. Words that fall into that category are replaced with the tag $<$ UNK $>$ in the Web1T corpus. Removing $<$ UNK $>$ words from the MPH has no effect on the counts in tier3 and is a straightforward process.

**Choosing N-Grams** N-grams can be as small as a 1-gram and grow, theoretically, to any size N imaginable. The preferred reference set for this thesis, the Web1T corpus, uses 1, 2, 3, 4, and 5-grams. While it is tempting to test all 5 N-gram sizes available in the corpus, only three were used. 1-grams and 5-grams were chosen to represent opposite ends of the size N gram spectrum available. 2-grams were used as a strong comparison to gappy bigrams and orthogonal sparse bigrams discussed below. Future work could focus on 3 and 4-grams to determine if there is a performance to size advantage in using those size of N-grams.

**Gappy Bigram and Orthogonal Sparse Bigram Construction** Once the 3 tier structure is created and functional, there are still two type of features remaining to be created. The Web1T corpus only contains standard n-grams, not gappy bigrams or orthogonal sparse bigrams. To create these more exotic types of bigrams, a rule for counting distance and a notation scheme was needed. It was decided to use ”lesser included counts” for both the gappy bigrams and the orthogonal sparse bigrams. This means that a word1 word2 pair would count for osb-0, osb-1, osb2, etc. While previous papers placed the distance for an OSB between word1 and word2 [19], this thesis constructed the OSBs with the distance after word2 for easier parsing. The gappy bigrams and OSBs were constructed from the 2, 3, 4, and 5-grams in the Web1T Corpus. Word pairs from a distance of 0 (a traditional bigram or an OSB-0) to a distance of 3 (an OSB-3 or the first and last word in a 5-gram) were built from the Web1T corpus. This process only looks at the first and last words in a 3-gram, 4-gram, or 5-gram since the inner words of this gram are already captured in the 2-gram. Using the inner 2-grams would double count 2-grams and throw off the count mass. The same is true for 3-grams inside of 4 and 5-grams as well as 4-grams inside of 5-grams.

**Grouping By Size** With references built and sized, an efficient structuring of the authors and documents needs to be devised. During data file construction, the grouping and conversion processes happened simultaneously. The grouping sets built were: small-to-large, small-and-large, and random.

**Small-To-Large** The small-to-large group matched the least prolific authors together with increasing size up to the most prolific authors. For example, of the 5 authors in the ENRON corpus with 5 total kilobytes worth of text are group together while the 5 authors with greater than 1 total megabyte of text are group together. No author is picked more than once. An example is shown in Figure 3.5.

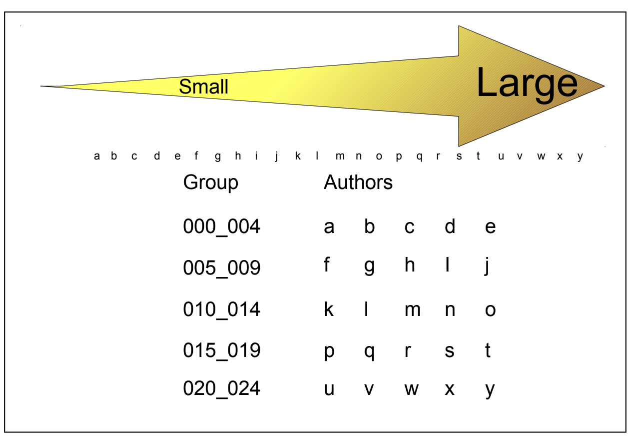
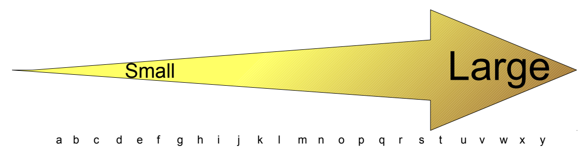



Figure 3.5: Small-To-Large Group for Group Size 5, 25 Authors

**Small-And-Large** The next group, small-and-large, is created by binning the authors by size. Then one author from each bin is picked to be group with one author from each other bin. For example the least prolific author is paired with one author from the most prolific bin and one author from each bin in between. In this situation, the selection from each bin is not random. The least prolific remaining author from each bin is picked for grouping. No author is picked more than once. An example is shown in Figure 3.6.




| Group | Authors | | | | |
|---|---|---|---|---|---|
| 000_004 | a | f | k | p | u |
| 005_009 | b | g | l | q | v |
| 010_014 | c | h | m | r | w |
| 015_019 | d | i | n | s | x |
| 020_024 | e | j | o | t | y |

Figure 3.6: Small-And-Large Group for Group Size 5, 25 Authors

**Random** This grouping simply produces a random number in the range of available authors and places the selected author into a group until that group is full. Then the next group is filled the same way until no authors remain. No author is picked more than once. No author is picked more than once. An example is shown in Figure 3.7.

**Group Sizes** Based on having 150 authors in the ENRON Corpus, the six following group sizes were used: 5, 10, 25, 50, 75 150. These six group sizes coupled with the three grouping types, small-to-large, small-and-large, and random creates 18 grouping types. Examples of these grouping types are 5 small-to-large, 5 small-and-large, 5 random, 10 small-to-large, ..., 150 small-to-large, 150 random. Although using all 150 authors in a grouping set makes the procedure of how the 150 were grouped redundant, all three size 150 tests were conducted as a check on the experiments. If the 150 author grouping provides different reslts, then there may be an issue with the classifiers.

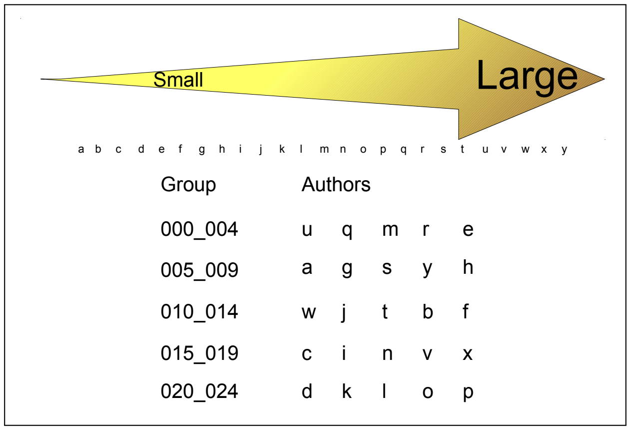



Figure 3.7: Random Group for Group Size 5, 25 Authors

After these grouping types were constructed, there were 171 totals sets (30 sets of 5 small-to-large, 15 sets of 10 small-to-large, ..., 1 set of 150 small-to-large, 1 set of 150 random.) Each of these sets were intended to be run through Bootstrapped SVM, Web1T SVM, Laplace Smoothed Naive Bayes, and Web1T Smoothed Naive Bayes. Assuming that only one MPH model is chosen to represent Google Web1T, that results in 684 experiments. Since there are 16 different MPH models based on the combinations punctuation, capitalization, sentence boundaries, and unknown words, the number of experiments could rise drastically. However, only two MPH models will be used during the experiments resulting in only 1,368 per feature type. Using 1-grams, 2-grams, 5-grams, 3-gb, and 3-osb results in 6,840 totals experiments.

**Data File Format** With combinations of features, artifacts, and group sizes chosen and the MPH data structures created, the actual documents must be converted into a format that can be used by the classifiers. The LibSVM file format was used since that it is the native format for LibLinear, the tool used for SVM in this thesis. The Naive Bayes classifier was built specifically for this thesis and was designed to use LibSVM format for convenience. The format of the data files consisted of an integer representing the author followed by a space, followed by a number representing the MPH value, followed by a colon, followed by another number representing the count. Each succeeding instance of a MPH value coupled with a count is separated by a space. Each document in the corpus is represented by a single line. Each line's mph number is in increasing order from left to right. The data files store the word/count pairs in a sparse fashion. This means that a zero count is not included in the data file. Absence of a word/count pair constitutes a zero count without needlessly using up space in the file. An example of this file format is provide in Figure 3.8.

```
83 362112:1 2216672:1 4609969:1 5582887:1 6141348:1 13588391:0
115 2334923:1 4077269:1 4759253:1 10878308:1 13069356:1 13588391:0
47 902626:1 1820755:1 10686459:1 12596717:1 13588391:0
80 1648944:1 1979998:1 2205090:1 2334923:1 2478205:2 13588391:0
```

Figure 3.8: LibSVM File Format

**Running SVM** With the data files created, the classifiers can be applied. The chosen tool for author detection using SVM is LibLinear. LibLinear was chosen for its speed compared to LibSVM. The LibLinear source code was slightly modified to allow training a model from a data set, then running prediction on a separate set without using the built-in cross validation function. During the training phase, each author has a SVM model built for it from a training file in a directory labeled ”train”. During the prediction phase, document contained in another file are used to predict the mostly likely author. That file is contained in a folder called ”predict”. The SVM author result is printed to a result file in a directory labeled ”result”. The f-score, precision, and recall for each file is recorded in a file inside a folder labeled ”analysis”. The analysis file also contains a full confusion matrix, time of prediction, size of original file, and other statistics. This file is finally pulled into a mySQL database for storage and calculation of precision, recall, and f-score..

The size of the author models impacts RAM usage and disk space. LibLinear stores SVM models as an array. RAM and storage are not the only limits. An array of integers representing token counts can be sizable, especially when token counts are long numbers (64 bits) instead of integers (32 bits).

RAM and disk storage are not the only limits. By specification, arrays in Java are limited to $2^{31} - 1$ entries. This means the model cannot contain more than $2^{31} - 1$ features. Also, the model must be loaded into RAM, so the number of authors coupled with the size of the author model must be weighed against the available RAM and disk storage.

**Running Naive Bayes** The Naive Bayes classifier has been specifically built for this thesis. The classifier reads in a pre-built array of long values from a file. The two types of arrays are a Laplace Smoothing array, which is comprised of all 1’s. the second type of array is the Google Smoothing array comprised of the count values from the Web1T corpus. Using an array to hold the smoothing values for Naive Bayes has an impact on RAM usage. There must be enough available RAM to hold the smoothing array. To prevent having numerous copies of the smoothing array in memory (one for each author being trained) a hashmap is used to create the author models instead. The process for training put each encountered feature type into a hashmap along with a count of $1 plus$the array smoothing value. If that feature type is encountered, the the count is simply incremented. Once all the training documents have been read and counted, the hashmaps of feature types and counts is converted into a hashmap of feature types and log of probability.

During the prediction process, each encountered feature type is queried against the author hashmap first. If the feature type is found in the hashmap, then the hashmap $\log probability$ is used. If not, then the smoothing array containing log of probabilities is used. An example of this hashmap/array process is shown in Figure 3.9. The result of the prediction process is outputted to a file in the corresponding results directory. Those results are then processed into a file in the corresponding analysis folder where all data is then read into a mySQL database for evaluation of precision, recall, and f-score.

## 3.3 Phase Two: Android Implementation

To manage files on the mobile device, a rudimentary file manager was built with a text viewer added. A button was also added to the File Manager to execute prediction against a document on the phone. An Android Service was also constructed that listens for incoming SMS messages. When an SMS Message is ”heard”, it is processed for author detection. The Service can be turned on and off using a button on the File Manager.

To measure CPU and RAM impact caused by the author detection processing, the third party applications, and Memory Usage, was installed on the phones. The method is to take a baseline of the phone’s CPU and RAM usage with no Widgets or Applications running, the phone is attached to a recharging device, and no calls or texts are being sent. The same phone conditions are being set for the processing tests where the only application that will run on the phone will be the SMS capture and author detection application for this thesis. This will yield some basic metrics of author detection impact on the phone’s capabilities.

## 3.4 Corpora

Two corpora are used for this thesis: the ENRON Email Corpus and the Naval Postgraduate School (NPS) Twitter Corpus. The aim of this thesis is to examine author detection using a mobile device. Two of the most common text communications on a mobile device are email

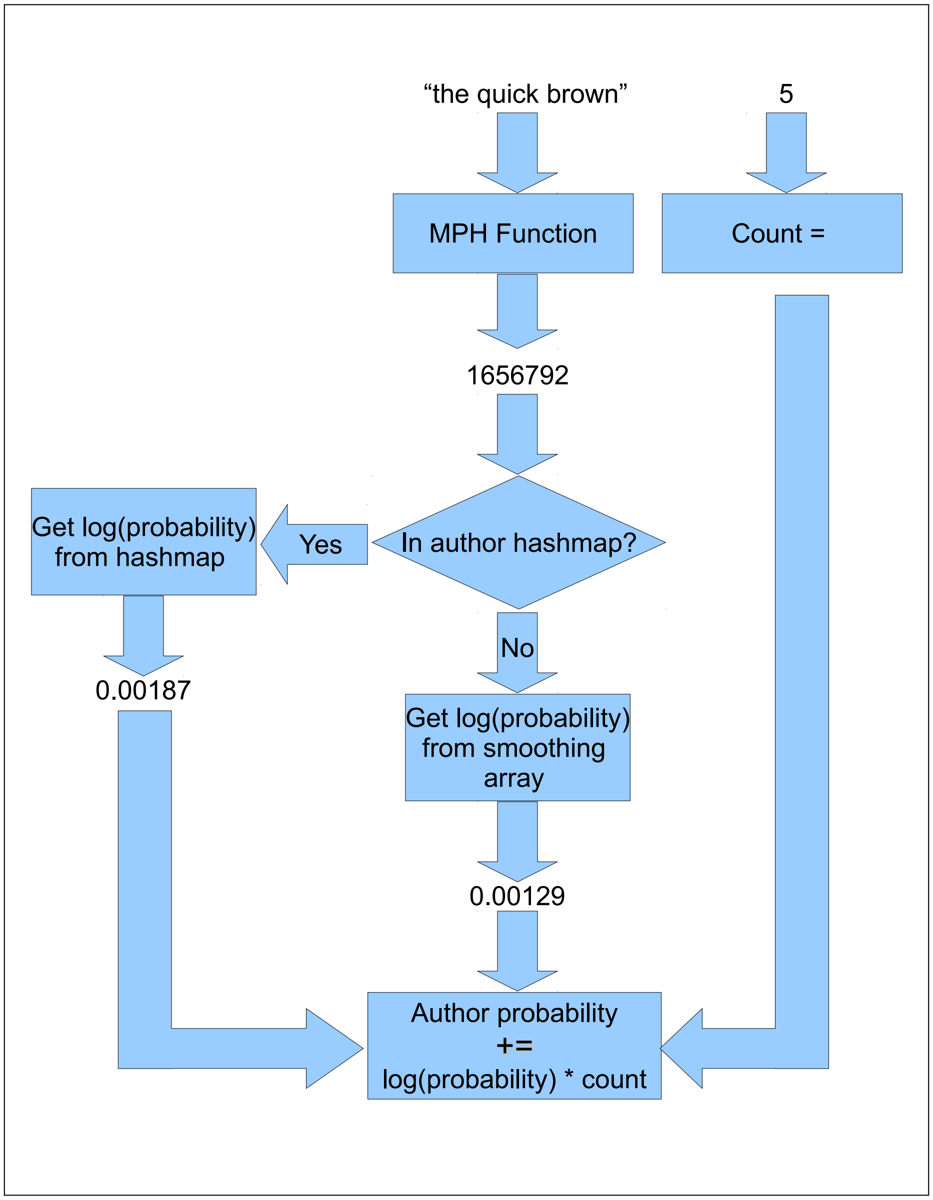



Figure 3.9: Naive Bayes Hashmap and Smoothing Array Flow Chart

and SMS (texting). The ENRON Email Corpus has been widely examined and has been used to author attribution in other studies. This makes the ENRON Corpus a suitable standard to measure the author detection techniques used in this thesis. The NPS Twitter Corpus is smaller and newer than the ENRON email corpus, but texting is extremely popular as a communications medium. Determining the effectiveness of author detection over this rapidly expanding text standard is important for analyzing the effectiveness of author detection on mobile devices.

**ENRON Email Corpus** Each ENRON email was stored in a single text file within a folder labeled with the author's first initial, second initial, and last name. Prior to processing each ENRON email, a systematic attempt was made to distill each email down into just the author's words. To support this distillation, the email header was stripped from each email. A search was conducted throughout the remaining text to find additional email headers. These are the embedded headers caused by email replies and forwards. Also to prevent biasing the author attribution, an attempt was made to systematically detect an email closing such as ”Sincerely, Dave” or ”Yours Truly, Jane”.

**Naval Postgraduate School Twitter Short Message Corpus** All tweets from a single author were stored in a single text file. Each tweet from that author was contained on its own line. Each line begins with a date-time stamp with the content of the text following. Prior to constructing the corpus, all ”re-tweets” were removed to ensure the text came from a single author, not just from a single Twitter account.

## 3.5 Intended Comparison

Once all tests are complete, performance of the different combinations of feature and classifiers will be compared for both the ENRON email corpus and the Twitter Corpus. This is to allow any differences in performance against the two primary media used on mobile phones. The completed test results should provide insight into the possibility of author detection on a mobile phone against both email and short messages.

# CHAPTER 4:
# Results and Analysis

## 4.1 Most Effective Combination of Classification Methods, Feature Types, and Vocabulary

Two measurements of effectiveness were used in this thesis: accuracy and f-score. Since the accuracy for each author is not the focus of this thesis, but rather the overall effectiveness of each classifier, feature type, and vocabulary combination, f-score is averaged for each combination. In each test set, average accuracy was higher then MLE. Likewise, average f-score was always lower than average accuracy.

At this point, it would be natural to simply compare the highest accuracy for each method-feature-vocabulary combination in the thesis and determine which combination performed best. This analysis would be flawed. Due to the underlying data structure in the Liblinear model, there is an absolute maximum number, $2^{31}$ of elements allowed in the model. There is one element created in the model for each feature-classifier combination. This means that the number of authors multiplied by the number of features cannot exceed $2^{31}$. Figure 4.1 shows the value of each feature-vocabulary-group combination. Cells highlighted in red cannot be used with the LibLinear model. There will be no LibLinear results for these combinations.

The impact of this hard maximum is large vocabularies show a higher accuracy and f-score than smaller vocabularies. This is not necessarily because the large vocabularies are more effective, but because the the larger vocabularies do not have the lower accuracy and f-score outcomes of the large group sizes. To illustrate this, the top ten feature-method combinations are show in Table 4.1 for the ENRON Email Corpus. The performance of each Liblinear OSB3-vocabulary combination is shown in Figure 4.2. Using Table 4.2 to evaluate effective would lead to a conclusion that LibLinear OSB3 has the best accuracy and f-score in this thesis. However, plotting all OSB3 results for each Web1t % in 4.2 shows that all OSB3-vocabulary combinations perform along a similar curve. The Web1T % $= 0$ is actually able to perform against all group sizes (5, 10, 25, 50, 75, and 150) and, thus, appears to perform worse than other OSB3s in the table, but clearly performs similarly from Figure 4.2. From this example, it becomes clear that simply using the table values in Appendix A through Appendix D provides an insufficient

| Feature Type | Web1T % | Liblinear Limits Due to Vocabulary Size (Web1T %) and Group Size | | | | | |
|---|---|---|---|---|---|---|---|
| | | Group Size | | | | | |
| | | 5 | 10 | 25 | 50 | 75 | 150 |
| GM1 | 1 | 679415 | 1358830 | 3397075 | 6794150 | 10191225 | 20382450 |
| | 2 | 1358835 | 2717670 | 6794175 | 13588350 | 20382525 | 40765050 |
| | 4 | 2717675 | 5435350 | 13588375 | 27176750 | 40765125 | 81530250 |
| | 8 | 5435355 | 10870710 | 27176775 | 54353550 | 81530325 | 163060650 |
| | 16 | 10870710 | 21741420 | 54353550 | 108707100 | 163060650 | 326121300 |
| GM2 | 1 | 15488310 | 30976620 | 77441550 | 154883100 | 232324650 | 464649300 |
| | 2 | 30976620 | 61953240 | 154883100 | 309766200 | 464649300 | 929298600 |
| | 4 | 61953240 | 123906480 | 309766200 | 619532400 | 929298600 | 1858597200 |
| | 8 | 123906480 | 247812960 | 619532400 | 1239064800 | 1858597200 | 3717194400 |
| | 16 | 247812960 | 495625920 | 1239064800 | 2478129600 | 3717194400 | 7434388800 |
| GM5 | 1 | 57357075 | 114714150 | 286785375 | 573570750 | 860356125 | 1720712250 |
| | 2 | 114714155 | 229428310 | 573570775 | 1147141550 | 1720712325 | 3441424650 |
| | 4 | 229428310 | 458856620 | 1147141550 | 2294283100 | 3441424650 | 6882849300 |
| | 8 | 458856620 | 917713240 | 2294283100 | 4588566200 | 6882849300 | 13765698600 |
| | 16 | 917713245 | 1835426490 | 4588566225 | 9177132450 | 13765698675 | 27531397350 |
| GB3 | 1 | 30275425 | 60550850 | 151377125 | 302754250 | 454131375 | 908262750 |
| | 2 | 60550850 | 121101700 | 302754250 | 605508500 | 908262750 | 1816525500 |
| | 4 | 121101700 | 242203400 | 605508500 | 1211017000 | 1816525500 | 3633051000 |
| | 8 | 242203405 | 484406810 | 1211017025 | 2422034050 | 3633051075 | 7266102150 |
| | 16 | 484406810 | 968813620 | 2422034050 | 4844068100 | 7266102150 | 14532204300 |
| OSB3 | 1 | 117215100 | 234430200 | 586075500 | 1172151000 | 1758226500 | 3516453000 |
| | 2 | 234430200 | 468860400 | 1172151000 | 2344302000 | 3516453000 | 7032906000 |
| | 4 | 468860400 | 937720800 | 2344302000 | 4688604000 | 7032906000 | 14065812000 |
| | 8 | 937720805 | 1875441610 | 4688604025 | 9377208050 | 14065812075 | 28131624150 |
| | 16 | 1875441615 | 3750883230 | 9377208075 | 18754416150 | 28131624225 | 56263248450 |

Figure 4.1: Liblinear Limits Due to Vocabulary Size and Group Size

analysis. A better analysis is provided by examining the plots in Appendix Q through Appendix T.

It is important to note that this is not an issue for combinations using Naive Bayes as a classification method. However, Naive Bayes did not outperform Liblinear in these tests, so a careful analysis of LibLinear using the plots in Appendix Q through Appendix T is required.

By examining the plots in Appendix Q through Appendix T, a clear trend emerges that the bootstrapped models, meaning models that made no use of the Web1T corpus as a vocabulary reference) performed similarly for liblinear to Web1T vocabularies. In all cases, the bootstrapped liblinear tests are usable for all group sizes. In this case, a good comparison would

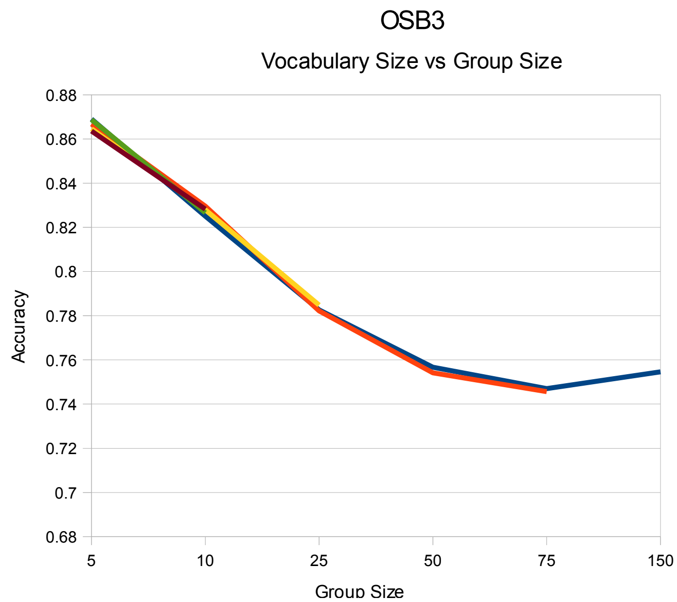



Figure 4.2: Accuracy of LibLinear OSB3 for the ENRON Email Corpus

be to drop all liblinear combinations that are not usable for all group sizes, then compare these remaining liblinear tests against all Naive Bayes tests. Since all Naive Bayes tests were usable for all group sizes, this makes the comparison fair.

After extracting out liblinear tests that were not usable against all groups sizes, the highest accuracy method-feature combination show the most accurate results for the ENRON Email Corpus in Table 4.1 below. The highest accuracy method-feature combination show the most accurate results for the Twitter Short Message Corpus in Table 4.2 below.

From Table 4.1 orthogonal sparse bigrams and gappy bigrams perform very well overall, with a traditional bigram making an entry at number five. The best performing method-feature combination is liblinear OSB3 with a Web1t% of 0. The next three combinations are Naive Bayes classifiers using OSB3 with large Web1T% vocabulary sizes. The results are similar for gappy bigrams, but at a reduced accuracy of approximately one percent.

From Table 4.2, the top performing method-feature combination is Naive Bayes OSB3 with a Web1T % of 0. The next four positions are filled with gappy bigrams with sizable Web1T% vocabularies. Why Twitter responds better to Naive Bayes as opposed to email responding better to liblinear is left to future work.

| Combinations | | Accuracy | | | | |
|---|---|---|---|---|---|---|
| Method | Feature Type | Web1T % | AVG | MIN | MAX | STDDEV |
| liblinear | OSB3 | 0 | 0.8362 | 0.5106 | 0.9732 | 0.1043 |
| nb | OSB3 | 16 | 0.8325 | 0.5213 | 0.9823 | 0.0890 |
| nb | OSB3 | 8 | 0.8315 | 0.5213 | 0.9714 | 0.0893 |
| nb | OSB3 | 4 | 0.8274 | 0.5197 | 0.9587 | 0.0924 |
| liblinear | GM2 | 0 | 0.8262 | 0.4824 | 0.9753 | 0.1087 |
| liblinear | GB3 | 0 | 0.8212 | 0.4787 | 0.9835 | 0.1121 |
| nb | GB3 | 16 | 0.8195 | 0.5201 | 0.9674 | 0.0947 |
| nb | GB3 | 4 | 0.8194 | 0.5340 | 0.9522 | 0.0941 |
| liblinear | GB3 | 1 | 0.8191 | 0.4731 | 0.9673 | 0.1110 |
| liblinear | GB3 | 2 | 0.8184 | 0.4765 | 0.9805 | 0.1113 |
| nb | GB3 | 8 | 0.8172 | 0.5255 | 0.9782 | 0.0935 |
| nb | OSB3 | 1 | 0.8126 | 0.3615 | 0.9574 | 0.1185 |
| nb | OSB3 | 2 | 0.8095 | 0.3526 | 0.9575 | 0.1283 |
| nb | OSB3 | 0 | 0.8058 | 0.5185 | 0.9592 | 0.0970 |
| liblinear | GM5 | 16 | 0.7918 | 0.3908 | 0.9676 | 0.1204 |
| liblinear | GM5 | 8 | 0.7872 | 0.3908 | 0.9513 | 0.1193 |
| nb | GB3 | 2 | 0.7857 | 0.4790 | 0.9669 | 0.1166 |
| liblinear | GM5 | 4 | 0.7755 | 0.3908 | 0.9455 | 0.1241 |
| liblinear | GM1 | 4 | 0.7742 | 0.4006 | 0.9590 | 0.1212 |
| liblinear | GM1 | 8 | 0.7740 | 0.4074 | 0.9570 | 0.1223 |
| liblinear | GM1 | 0 | 0.7735 | 0.3776 | 0.9531 | 0.1222 |

Table 4.1: Highest Accuracy Method-Feature Type Combinations for the ENRON Email Corpus

While the above table shows the best performing, accuracy is not always a solid measure of classification effectiveness. A better measure is f-score. As shown repeatedly by the tables in Appendix A through Appendix D, the relative performance of average f-score matched the relative performance of accuracy for each test set. In all cases, f-score was lower than the average accuracy. Even more telling about the results is every test set shows a minimum f-score of 0. That means that at least one author had an f-score of zero in each test. This accounts for the high standard deviation for f-scores across all tests. For f-scores of approximately 0.65 the standard deviation was approximately 0.25.

An examination of the confusion matrices for each test can provide insight into whether there was a ”poison” author that never got selected or if there was an author who was a selection ”magnet” always getting too many selections for documents. Due to the large number of confusions matrices in this thesis ( nearly 19,782 confusion matrices created from 57 tests * 3 size

| Combinations | | Accuracy | | | | |
|---|---|---|---|---|---|---|
| Method | Feature Type | Web1t % | AVG | MIN | MAX | STDDEV |
| nb | OSB3 | 0 | 0.5525 | 0.2320 | 0.8164 | 0.1339 |
| nb | GB3 | 16 | 0.5327 | 0.2216 | 0.8216 | 0.1351 |
| nb | GB3 | 4 | 0.5271 | 0.2190 | 0.8546 | 0.1375 |
| nb | GB3 | 8 | 0.5256 | 0.2176 | 0.8474 | 0.1362 |
| nb | GB3 | 2 | 0.5249 | 0.2186 | 0.7823 | 0.1324 |
| liblinear | GM2 | 4 | 0.5228 | 0.1809 | 0.8210 | 0.1477 |
| nb | GB3 | 1 | 0.5204 | 0.2148 | 0.8125 | 0.1319 |
| nb | GB3 | 0 | 0.5203 | 0.1973 | 0.8021 | 0.1389 |
| liblinear | GM2 | 1 | 0.5197 | 0.1882 | 0.8454 | 0.1483 |
| liblinear | GM1 | 8 | 0.5187 | 0.1743 | 0.9026 | 0.1525 |
| liblinear | GM2 | 2 | 0.5186 | 0.1830 | 0.8232 | 0.1495 |
| liblinear | GM1 | 1 | 0.5159 | 0.1768 | 0.8211 | 0.1494 |
| liblinear | GM1 | 4 | 0.5149 | 0.1874 | 0.8546 | 0.1485 |
| liblinear | GM1 | 0 | 0.5141 | 0.1802 | 0.8089 | 0.1485 |
| nb | GM1 | 0 | 0.5140 | 0.1247 | 0.7714 | 0.1631 |
| liblinear | GM1 | 16 | 0.5134 | 0.1865 | 0.8324 | 0.1483 |
| liblinear | GM1 | 2 | 0.5131 | 0.1818 | 0.8966 | 0.1487 |
| liblinear | GM5 | 1 | 0.4768 | 0.1398 | 0.8362 | 0.1521 |
| nb | GM2 | 0 | 0.4750 | 0.1630 | 0.7890 | 0.1406 |
| nb | OSB3 | 2 | 0.4739 | 0.1790 | 0.7734 | 0.1370 |
| nb | OSB3 | 8 | 0.4707 | 0.1787 | 0.7790 | 0.1373 |

Table 4.2: Highest Accuracy Method-Feature Type Combinations for the Twitter Short Message Corpus

groupings * 6 vocabulary sizes * 5 feature types * 2 corpora * 2 methods - 738 unusable LibLinear tests) the confusions matrices are not presented in this thesis, but are archived by the NPS Natural Language Processing lab in comma separated value files.

## 4.2 Impact of Author Relative Prolificity on Classifier Effectiveness

While identifying the best accuracy results for method-feature combinations is important, these results could mask a weakness in the method-feature combinations. Does the relative prolificity of each author drastically impact the results? To answer this question, the tests in this thesis were conducted in three groupings: small-to-large, small-and-large, and random. As explained fully in Chapter 3, these groupings were based on a rank-ordering by size for each author's total document collection. For small-to-large, the least prolific authors are grouped together, the most prolific authors are grouped together. The idea behind the small-to-large group is to keep the difference in size between the authors to a minimum. For small-and-large, the opposite idea is employed. The smallest authors are combined with the largest authors using a bucket strategy. Each bucket contains rank-ordered by size authors of similar size. One author is picked from each bucket to provide a maximum variety of author document collections sizes. In the random group, the authors are grouped together using a pseudo-random number generator, where each author has been assigned a number.

The results of testing in this thesis for accuracy and f-score, broken out by small-to-large, small-and-large, and random are given in Appendix E through Appendix H. The results from Appendix E, LibLinear Results for the ENRON Email Corpus, show the accuracy for small-to-large is always lower than the accuracy for small-and-large and random. However, the f-score for small-to-large is always higher than the f-score for small-and-large and random. This result shows how accuracy is dominated by the MLE author, since allowing a more prolific author into a group with less prolific authors tends to raise accuracy, but hurts f-score. To illustrate the effect of author prolificity on accuracy and f-score Table 4.3 shows the confusion matrix for a small-to-large grouping of size 10 for GB3, Web1T%=0. Table 4.4 shows the confusion matrix for a small-and-large grouping of size 10 for GB3, Web1T%=0.

| | | Label | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 11 | 111 | 119 | 14 | 146 | 15 | 48 | 60 | 71 | 91 |
| Truth | 11 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 0 |
| | 111 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 119 | 0 | 0 | 8 | 1 | 0 | 0 | 0 | 0 | 0 | 6 |
| | 14 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 10 |
| | 146 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| | 15 | 0 | 0 | 0 | 1 | 0 | 4 | 1 | 0 | 0 | 4 |
| | 48 | 0 | 0 | 0 | 2 | 0 | 0 | 9 | 0 | 0 | 2 |
| | 60 | 0 | 0 | 2 | 0 | 0 | 0 | 1 | 4 | 0 | 2 |
| | 71 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 4 |
| | 91 | 0 | 0 | 0 | 2 | 0 | 0 | 1 | 0 | 0 | 17 |

Table 4.3: Confusion Matrix for Small-To-Large Grouping, Feature Type: GB3, Group Size: 10, Web1T%: 0

Table 4.3 represents a group of similarly prolfic authors. One author, author 91, not only has the highest number of true positives, 17, but has a large number of false positives. The combined false positives for all other authors is 21, compared to author 91's 29 false postives. That counts as 29 false negatives spread across the other 9 authors, impacting their false negative value. For

| | | Label | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 11 | 113 | 47 | 49 | 58 | 75 | 76 | 86 | 88 | 95 |
| Truth | 11 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 2 | 0 |
| | 113 | 0 | 203 | 43 | 23 | 3 | 6 | 0 | 4 | 19 | 0 |
| | 47 | 0 | 7 | 2510 | 2 | 4 | 2 | 0 | 3 | 61 | 0 |
| | 49 | 0 | 16 | 52 | 1180 | 2 | 6 | 0 | 2 | 48 | 1 |
| | 58 | 0 | 1 | 16 | 2 | 508 | 0 | 0 | 0 | 7 | 0 |
| | 75 | 0 | 5 | 19 | 4 | 0 | 338 | 0 | 1 | 16 | 0 |
| | 76 | 0 | 0 | 1 | 3 | 0 | 0 | 9 | 0 | 1 | 0 |
| | 86 | 0 | 14 | 12 | 14 | 2 | 9 | 0 | 36 | 15 | 0 |
| | 88 | 0 | 11 | 129 | 12 | 1 | 7 | 0 | 0 | 277 | 1 |
| | 95 | 0 | 4 | 2 | 7 | 3 | 2 | 0 | 1 | 9 | 4 |

Table 4.4: Confusion Matrix for Small-And-Large Grouping, Feature Type: GB3, Group Size: 10, Web1T%: 0

calculating f-score, a higher false negative rate decreases recall and, since true postives remain constant, false postives fall, increasing precision. In the small-to-large grouping, one author has very few false postives, creating a high precision. The other authors end up with a high recall. As the f-score for each author is average for the group, these unbalanced numbers drive the f-score higher while maintaining a lower accuracy.

Table 4.4 represents a group of dissimilarly prolific authors. In this grouping, one author does not dominate the number of false positives. This more evenly spread set of false positives and false negatives keeps the overal f-score lower, while maintaining a higher accuracy. The bottom line is the high outlier precision score for one author in the small-to-large group gives a higher f-score, but lower accuracy. A median measurement of f-score might provide a better picture of overall f-score behavior than an average f-score.

The other issue that arises from the f-score average is the small-to-large f-score has a smaller standard deviation than the small-and-large f-score. This points to a tighter grouping of values. This arises from all but one author having similar f-score values. The small-and-large group has no single outlier f-score to drag the f-score higher, but the values do have greater variation among all points.

The above paragraphs make use of a cursory examination of the behavior of author detection due to author prolificity. An in depth statistical analysis of the difference between the author groupings is warranted as future work. The goal of using these different groupings was to ensure that the tools chosen in this thesis behaved predictably with respect to varying author prolificity within a detection group. To examine that behavior, plots of accuracy, average f-score, MLE, precision, and recall for each method-feature combination across all usable Web1T% vocabularies is included in Appendix Q thorugh Appendix U. To illustrate that the impact of author prolificity is predictable across method-feature combinations and corpora, Figure 4.3 and Figure 4.4 are show as representative samples of overall classifer and corpora results.

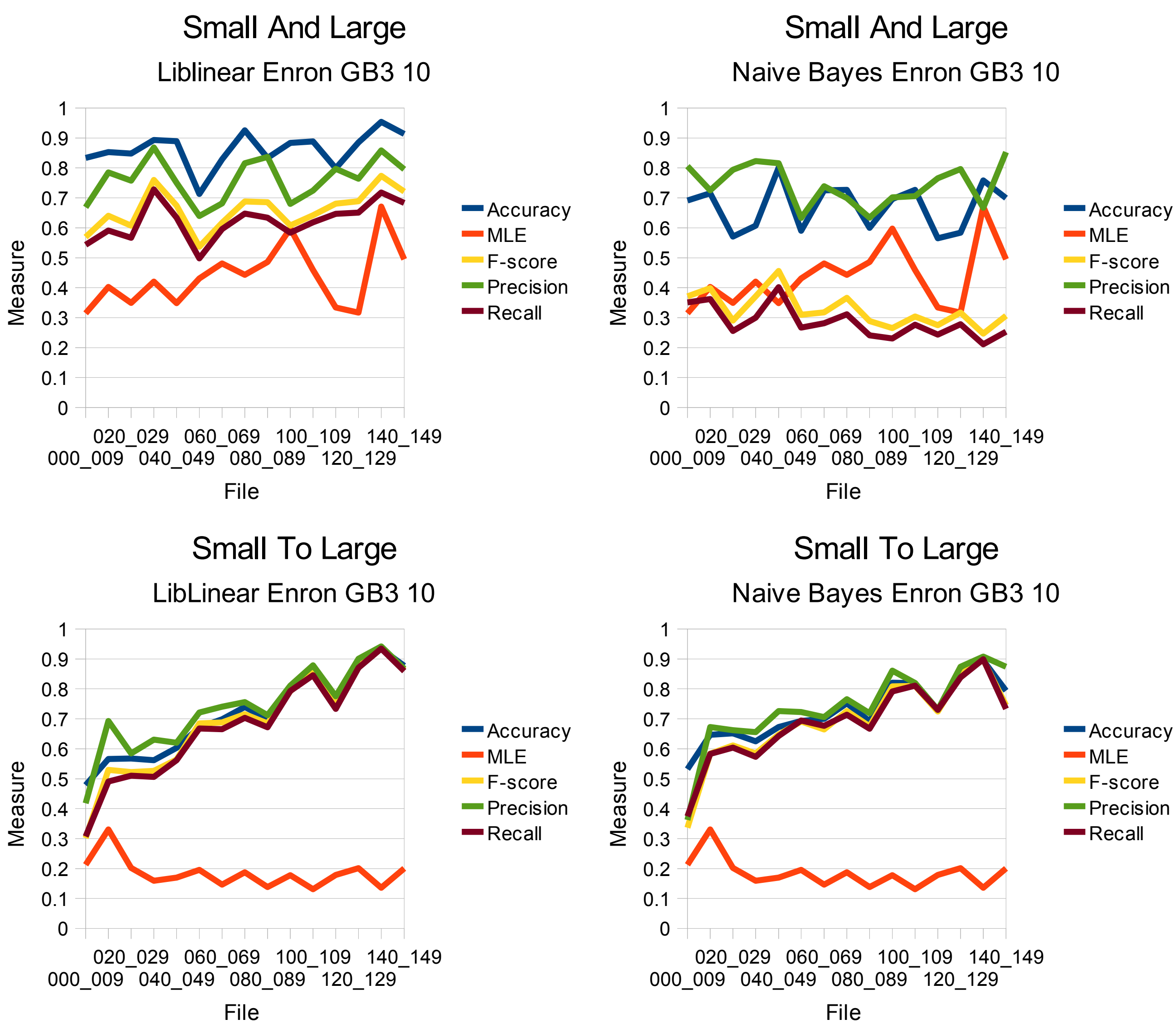



Figure 4.3: Liblinear Limits Due to Vocabulary Size and Group Size

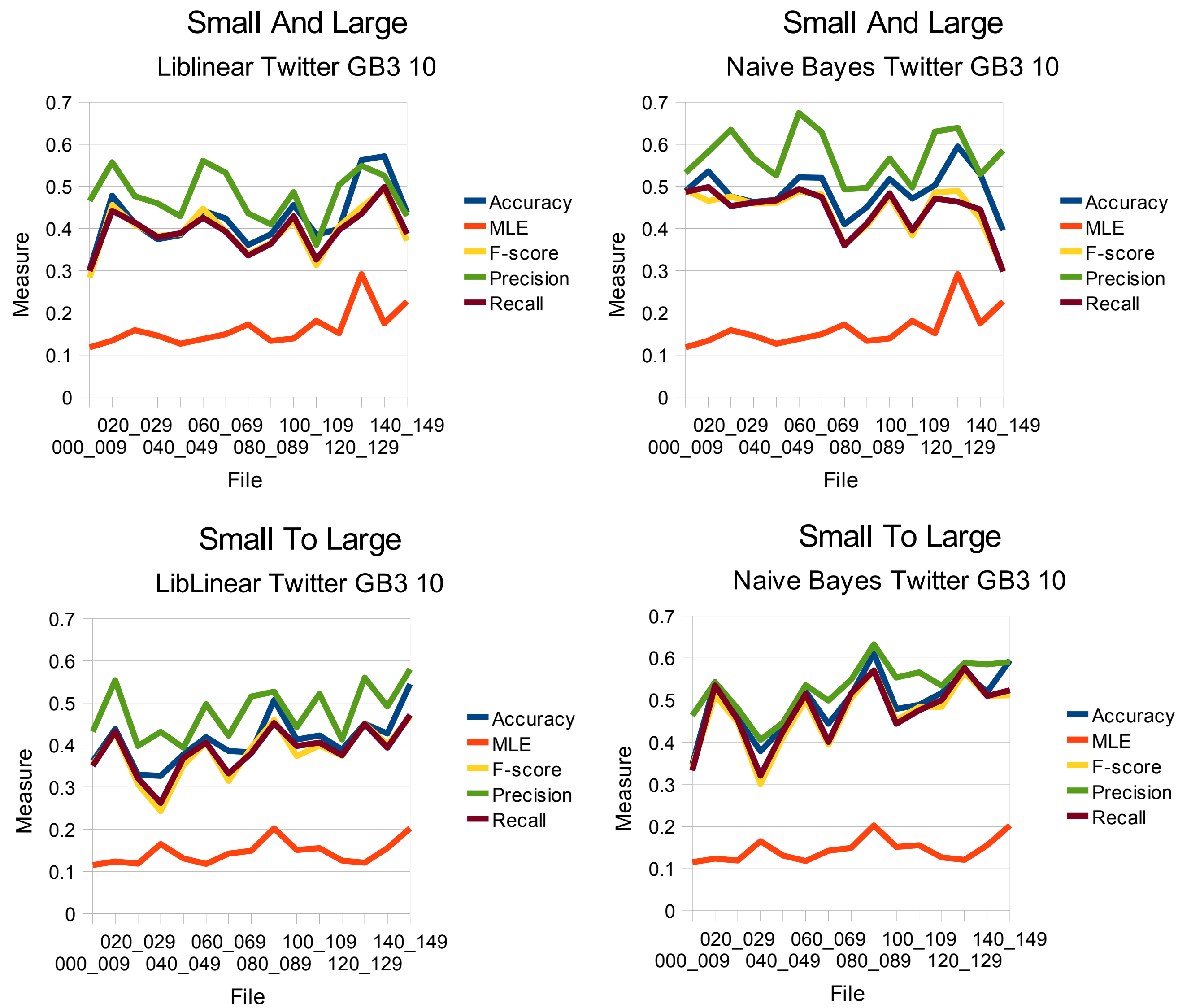



Figure 4.4: Liblinear Limits Due to Vocabulary Size and Group Size

From Figure 4.3 some trends become apparent. As the small-to-large graph for the Enron corpus moves from left to right, the accuracy, f-score, precision, and recall all increase in tight agreement. This correlates to the wide variation in proflificity between the least prolific group on the far left, file 000_009, and the last file on the far right, file 140_149. In the Enron corpus, the least prolific author's document total size is measured in a few kilobytes where the most prolfic author's document total size is measured in megabytes. Most striking is that the trend holds for both Liblinear and Naive Bayes. Also, with a group size of 10, the most prolific authors have a high accuracy, high f-score, high precision, and high recall. The impact of prolificity is predictable and significant for the Enron Corpus.

The results for the Enron corpus small-and-large group are largely flat as the graph moves from left to right. This shows that in a mixed group of varying prolificity, both liblinear and Naive Bayes maintain fairly consistent results. Clearly, having an author who is significantly more prolific than other authors in his detection group hurts the average f-score for that group while raising the accuracy. This raise is accuracy is not a good indicator of improved performance. For the Enron Email Corpus, proflific authors are more detectable than less prolific authors, even in the presence of other prolific authors.

In the Enron small-and-large figures, precision and accuracy are close in value where f-score and recall are always close in value. The accuracy and precision values are also always above the f-score and recall values. Investigation into the underlying reasons for this pattern warrants future work in an in-depth statistical analysis of the effects of grouping on author detection.

The story from Figure 4.4 is markedly different than the story from Figure 4.3. All results are lower for the Twitter Short Message Corpus than for the Enron Email Corpus, as indicated by the top value on for graphs in Figure 4.4 being 0.7 versus 1.0 for Figure 4.3. The relative flatness of measures in the Twitter corpus compared to the Enron corpus can be explained by the difference in relative sizes of an author in the Twitter corpus and an author in the Enron corpus. The most prolific author in the Twitter corpus has only 15.2KB of text as opposed to 2.5MB for the most prolific Enron author. Future work of gathering a larger Twitter corpus of original, not re-tweeted, short messages could supply a similiar size and variation of the Enron corpus.

## 4.3 Storage Requirements for Combinations of Classification Methods, Feature Types, and Vocabulary

While the effectiveness of the method-feature combinations are important, these tools are of no use on a mobile device unless the tool can actually fit on the mobile device. An important fact about determining the size of classifier models is that the size of the model in RAM does not equal the size of the model when written to a file. For instance, a Java long (primitive) of 1 uses 8 bytes of RAM, but is represented in a file using only 4 bytes. Similarly, there is a disparity between the UTF-8 values byte size on disk and the object representation in RAM for many

Java objects. This is why heap size could not be used as an accurate measurement of model size.

To determine if any of these method-feature combinations will fit on a mobile device, a few combinations had exhaustive outputs of their model sizes computed. After determining that the standard deviation for models with a vocabulary size greater than a Web1T% of 0 was trivial, only a small sample of the remaining method-feature combinations were computed. Due to the large size of many models, only one model size was calculated for many method-feature combinations.

Actually writing out these models to disk would have been extremely time consuming and a load on the already taxed Hamming High Performance Cluster. To conduct the size measurements, the liblinear models were written to a Java ByteArrayOutputStream. Once the write was complete, the size of the ByteArrayOutputStream buffer was measured. The worked well for models smaller than 2GB. Models larger than 2GB caused the ByteArrayOutputStream to be ”full” since the index for an ByteArrayOutputStream is limited to $2^{31}$ elements and each element in that array is a byte. For any model larger than 2GB, the size for that model was not recorded and thus has no size record in Appendix M through Appendix P nor a score in the scoring tables in Appendix I through Appendix L.

What constitutes a storage requirement for the method-feature combinations in this thesis depends on the vocabulary size and method used. A Web1T% of 0 in liblinear requires no keys.mph or signature file, but does require a sizable vocabulary map file to functions. For Naive Bayes, a Web1T% of 0 does not require a keys.mph file, signature file, count file, nor logprobs file. However a sizable vocabulary map is needed. The sizes for each combination’s keys.mph, signature, counts, logprobs, and average author size are included with totals in Appendix M through Appendix P. To provide an intuition on the magnitude of sizes involved, Table 4.5 shows sizes for keys.mph, signature, counts, logprobs, and vocabmap for a few method-feature combinations. Table 4.5 shows only the vocabmap size for the Web1T% of 0. This is because Web1T% of 0 does not use keys.mph, signature, counts, or logprobs references, but does create it own vocabulary map. Complete size tables are provided in Appendix M through Appendix P.

It is quickly apparent from this table that few of these files could be loaded into the RAM of a 16MB Dalvik VM. If these files were to be used, they would have to be read directly from the microSD card, which is an expensive operation compared to reading from RAM. A

| Method | Feature Type | Web1T% | Size (MB) | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | keys.mph | signature | counts | logprobs | vocabmap | Total |
| liblinear | GB3 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 54.31 | 54.31 |
| liblinear | GB3 | 1 | 3.21 | 12.11 | 0.00 | 0.00 | 0.00 | 15.32 |
| liblinear | GB3 | 2 | 6.41 | 24.22 | 0.00 | 0.00 | 0.00 | 30.63 |
| liblinear | GB3 | 4 | 12.82 | 48.44 | 0.00 | 0.00 | 0.00 | 61.27 |
| liblinear | GB3 | 8 | 25.64 | 96.89 | 0.00 | 0.00 | 0.00 | 122.53 |
| liblinear | GB3 | 16 | 51.31 | 193.85 | 0.00 | 0.00 | 0.00 | 245.15 |
| liblinear | GM1 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 1.40 | 1.40 |
| liblinear | GM1 | 1 | 0.07 | 0.27 | 0.00 | 0.00 | 0.00 | 0.34 |
| liblinear | GM1 | 2 | 0.14 | 0.54 | 0.00 | 0.00 | 0.00 | 0.69 |
| liblinear | GM1 | 4 | 0.29 | 1.09 | 0.00 | 0.00 | 0.00 | 1.37 |
| liblinear | GM1 | 8 | 0.58 | 2.17 | 0.00 | 0.00 | 0.00 | 2.75 |
| liblinear | GM1 | 16 | 1.15 | 4.35 | 0.00 | 0.00 | 0.00 | 5.50 |
| nb | GB3 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 54.31 | 54.31 |
| nb | GB3 | 1 | 3.21 | 12.11 | 48.44 | 48.44 | 0.00 | 112.20 |
| nb | GB3 | 2 | 6.41 | 24.22 | 96.88 | 96.88 | 0.00 | 224.39 |
| nb | GB3 | 4 | 12.82 | 48.44 | 193.78 | 193.78 | 0.00 | 448.83 |
| nb | GB3 | 8 | 25.64 | 96.89 | 387.55 | 387.55 | 0.00 | 897.64 |
| nb | GB3 | 16 | 51.31 | 193.85 | 775.39 | 775.39 | 0.00 | 1795.94 |
| nb | GM1 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 1.40 | 1.40 |
| nb | GM1 | 1 | 0.07 | 0.27 | 1.09 | 1.09 | 0.00 | 2.52 |
| nb | GM1 | 2 | 0.14 | 0.54 | 2.17 | 2.17 | 0.00 | 5.04 |
| nb | GM1 | 4 | 0.29 | 1.09 | 4.35 | 4.35 | 0.00 | 10.07 |
| nb | GM1 | 8 | 0.58 | 2.17 | 8.70 | 8.70 | 0.00 | 20.14 |
| nb | GM1 | 16 | 1.15 | 4.35 | 17.40 | 17.40 | 0.00 | 40.29 |

Table 4.5: Sample of Vocabulary Reference File Sizes

more thorough discussion of method-feature combinations is discussed in the last section of this chapter.

Apart from the vocabulary references needed for the method-feature combinations, each method-feature combination produces a different authors model size. Unlike the vocabulary reference files, the authors model file sizes vary greatly. The model constructed for liblinear consists of an array populated with the support vector values for each author. The model for Naive Bayes consists of a Java hashmap. That hashmap has an Integer object for a key and a Double object for its value. The Integer object is the mapped integer value for a given token. The Double object is the probability for that token during the training process.

The impact of authors model size for a mobile device is important. Even if the vocabulary reference files can be accomodated by a mobile device, a large authors model can push the storage requirement beyond the 16MB Dalvik VM's capability or even the capacity of common microSD cards. It is important to note here that size on a file only provides a relative indicator of size in RAM for a given method-feature combination. Actually measuring the impact of Dalvik VM in terms of RAM used versus storage requirements is left to future work as this study involves how model referencing is handled and how values on the file are converted to objects in memory. Table 4.6 shows a sample of author sizes for both liblinear and Naive Bayes authors models. A complete list of average authors models sizes is provided in Appendix M through Appendix P.

## 4.4 Classification Effectiveness Versus Storage Requirements

With the resource constraints of mobile devices and the author detection reqirements of this thesis, some method must be used to evaluate the tradeoff between accuracy and size. For this thesis, effectivness will be divided by the full storage requiment for each method-feature combination. The storage requirements will be computed as the sum of keys.mph, signature, counts, logprobs, vocabmap, and average authors model size for each method-feature combination. The complete set of scores for this thesis are included in Appendix I through Appendix L.

*It is important to note that there are no scores for any authors model size over 2GB. This is due to the limitations of measuring on-disk size for authors models with a ByteArrayOutputStream, but this limitation will not adversely affect the conclusions of this thesis. Any authors model larger than 2GB is impractical for current mobile devices. Also a 2GB divisor for the score computation would put that method-feature combination out of contention for a top performer in this thesis.

The top performing method-feature combination for the Enron Email Corpus was Naive Bayes method using GM1 for group size 5 with a score of 0.4495. Table 4.7 shows the top 20 scores along with accuracy and size information for the Enron Email Corpus. All of these top performers use the GM1 feature type. The accuracy of these combinations is in the same range as the most accurate method-feature combinations. However, these accuracies are mostly for group

| Corpus | Method | Feature Type | Group Size | Web1T% | Size (MB) AVG | MIN | MAX | STDDEV |
|---|---|---|---|---|---|---|---|---|
| enron | liblinear | OSB3 | 5 | 0 | 15.254 | 8.020 | 31.368 | 5.840 |
| enron | liblinear | OSB3 | 5 | 1 | 259.320 | 211.231 | 262.991 | 7.944 |
| enron | liblinear | OSB3 | 5 | 2 | 521.039 | 422.022 | 530.023 | 11.188 |
| enron | liblinear | OSB3 | 5 | 4 | 1031.477 | 844.102 | 1039.616 | 26.316 |
| enron | nb | OSB3 | 5 | 0 | 5.328 | 0.068 | 34.479 | 7.090 |
| enron | nb | OSB3 | 5 | 1 | 8.528 | 0.075 | 54.680 | 11.243 |
| enron | nb | OSB3 | 5 | 2 | 8.544 | 0.075 | 54.939 | 11.286 |
| enron | nb | OSB3 | 5 | 4 | 8.550 | 0.075 | 55.054 | 11.305 |
| enron | nb | OSB3 | 5 | 8 | 8.553 | 0.075 | 55.100 | 11.314 |
| enron | nb | OSB3 | 5 | 16 | 8.554 | 0.075 | 55.121 | 11.317 |
| twitter | liblinear | GM1 | 5 | 0 | 0.088 | 0.076 | 0.108 | 0.007 |
| twitter | liblinear | GM1 | 5 | 1 | 1.568 | 1.546 | 1.614 | 0.013 |
| twitter | liblinear | GM1 | 5 | 2 | 3.064 | 3.043 | 3.109 | 0.013 |
| twitter | liblinear | GM1 | 5 | 4 | 6.050 | 6.013 | 6.099 | 0.015 |
| twitter | liblinear | GM1 | 5 | 8 | 12.034 | 12.011 | 12.079 | 0.013 |
| twitter | liblinear | GM1 | 5 | 16 | 23.952 | 23.869 | 24.038 | 0.037 |
| twitter | nb | GM1 | 5 | 0 | 0.024 | 0.016 | 0.045 | 0.005 |
| twitter | nb | GM1 | 5 | 1 | 0.040 | 0.034 | 0.050 | 0.003 |
| twitter | nb | GM1 | 5 | 2 | 0.040 | 0.035 | 0.051 | 0.003 |
| twitter | nb | GM1 | 5 | 4 | 0.040 | 0.036 | 0.052 | 0.003 |
| twitter | nb | GM1 | 5 | 8 | 0.040 | 0.034 | 0.053 | 0.003 |
| twitter | nb | GM1 | 5 | 16 | 0.040 | 0.035 | 0.050 | 0.003 |

Table 4.6: Sample of Authors Model File Sizes

sizes of 5, 10, and 25, which limits the applicability of the tools in this thesis. There is only one combination for group size 50 and only one combination of group size 75. All of these top 20 scores have storage requirements under 16MB.

The top performing method-feature combination for the Twitter Short Message Corpus was Naive Bayes using feature type GM1 for a group size of 5. Table 4.8 shows the top 20 scores along with accuracy and size information for the Twitter Short Message Corpus. The accuracy of these combinations is in the same range as the most accurate method-feature combinations. The range of groups sizes that made the top 20 scores is much larger than for the ENRON Email

| Method | Corpus | Feature Type | Group Size | Web1T % | Score | Accuracy | Size(MB) |
|---|---|---|---|---|---|---|---|
| nb | enron | GM1 | 5 | 0 | 0.4495 | 0.7215 | 1.60 |
| liblinear | enron | GM1 | 5 | 0 | 0.4374 | 0.8269 | 1.89 |
| liblinear | enron | GM1 | 5 | 1 | 0.3685 | 0.8233 | 2.23 |
| nb | enron | GM1 | 10 | 0 | 0.3186 | 0.5768 | 1.81 |
| liblinear | enron | GM1 | 10 | 0 | 0.2789 | 0.7611 | 2.73 |
| nb | enron | GM1 | 5 | 1 | 0.2262 | 0.6441 | 2.85 |
| liblinear | enron | GM1 | 5 | 2 | 0.2017 | 0.8216 | 4.07 |
| liblinear | enron | GM1 | 10 | 1 | 0.1800 | 0.7610 | 4.23 |
| nb | enron | GM1 | 25 | 0 | 0.1683 | 0.4083 | 2.43 |
| nb | enron | GM1 | 10 | 1 | 0.1634 | 0.5189 | 3.18 |
| nb | enron | GM1 | 5 | 2 | 0.1212 | 0.6505 | 5.37 |
| liblinear | enron | GM1 | 25 | 0 | 0.1124 | 0.6845 | 6.09 |
| liblinear | enron | GM1 | 5 | 4 | 0.1071 | 0.8298 | 7.75 |
| liblinear | enron | GM1 | 10 | 2 | 0.1024 | 0.7594 | 7.42 |
| nb | enron | GM1 | 25 | 1 | 0.0950 | 0.3956 | 4.16 |
| nb | enron | GM1 | 10 | 2 | 0.0915 | 0.5215 | 5.70 |
| nb | enron | GM1 | 50 | 0 | 0.0903 | 0.3126 | 3.46 |
| liblinear | enron | GM1 | 25 | 1 | 0.0648 | 0.6847 | 10.56 |
| nb | enron | GM1 | 75 | 0 | 0.0648 | 0.2912 | 4.50 |
| nb | enron | GM1 | 5 | 4 | 0.0635 | 0.6610 | 10.40 |

Table 4.7: Highest Scoring Method-Feature Combinations for the Enron Email Corpus

Corpus. There are three combinations for group size 50, two combinations of group size 75, and two combinations of group size 150. All of the top 20 performing score combinations have a storage requirement of less than 16MB.

With the scores measure for each method-feature combination in hand, the shortcoming of using $score = \frac{accuracy}{size}$ become apparent. Table 4.7 indicates that Naive Bayes using GM1 for group size 5 is the best feature-combination to choose for a mobile device. However, the second highest score, liblinear using GM1 for group size 5 has an accuracy of 0.8269 where the top scoring combination has an accuracy of 0.7215, a full 0.1 worse than the second top scorer. An even more important limitation to this approach if the heavy bias of group size on the scoring process. To address this, Table 4.9 for the Enron Email Corpus and Table 4.10 for the Twitter

| Method | Corpus | Feature Type | Group Size | Web1T % | Score | Accuracy | Size(MB) |
|---|---|---|---|---|---|---|---|
| nb | twitter | GM1 | 5 | 0 | 2.8233 | 0.6264 | 0.22 |
| nb | twitter | GM1 | 10 | 0 | 1.9815 | 0.4869 | 0.25 |
| liblinear | twitter | GM1 | 5 | 0 | 2.1731 | 0.6212 | 0.29 |
| nb | twitter | GM1 | 25 | 0 | 1.0593 | 0.3357 | 0.32 |
| nb | twitter | GM1 | 50 | 0 | 0.5347 | 0.2326 | 0.43 |
| liblinear | twitter | GM1 | 10 | 0 | 1.0850 | 0.4762 | 0.44 |
| nb | twitter | GM1 | 75 | 0 | 0.3291 | 0.1820 | 0.55 |
| nb | twitter | GM1 | 150 | 0 | 0.1375 | 0.1252 | 0.91 |
| liblinear | twitter | GM1 | 25 | 0 | 0.3301 | 0.3461 | 1.05 |
| nb | twitter | GM2 | 5 | 0 | 0.4866 | 0.5711 | 1.17 |
| nb | twitter | GM2 | 10 | 0 | 0.3644 | 0.4439 | 1.22 |
| nb | twitter | GM2 | 25 | 0 | 0.2380 | 0.3215 | 1.35 |
| liblinear | twitter | GM2 | 5 | 0 | 0.3503 | 0.4844 | 1.38 |
| nb | twitter | GM2 | 50 | 0 | 0.1598 | 0.2509 | 1.57 |
| nb | twitter | GM2 | 75 | 0 | 0.1207 | 0.2162 | 1.79 |
| liblinear | twitter | GM1 | 5 | 1 | 0.3257 | 0.6228 | 1.91 |
| liblinear | twitter | GM2 | 10 | 0 | 0.1911 | 0.3700 | 1.94 |
| liblinear | twitter | GM1 | 50 | 0 | 0.1153 | 0.2693 | 2.34 |
| nb | twitter | GM2 | 150 | 0 | 0.0696 | 0.1709 | 2.46 |
| nb | twitter | GM1 | 5 | 1 | 0.1945 | 0.4974 | 2.56 |

Table 4.8: Highest Scoring Method-Feature Combinations for the Twitter Short Message Corpus

Short Message Corpus were constructed to show the score for each feature-method-percentage combination that could cover all group sizes with score averaged over all group sizes.

The top method-feature combinations in Table 4.9 are still dominated by GM1 as a feature type. Naive Bayes using GM1 and a Web1T%=0 had a higher score than Liblinear using GM1 and a Web1T%=0, but the liblinear accuracy is 0.1284 higher than the Naive Bayes accuracy. This shows again that this scoring method by itself does not produce an optimal feature-method combination on its own.

Similarly, the top method-feature combinations for the Twitter Short Message Corpus in Table 4.10 are GM1. However, there is a much wider mix of feature types in the Twitter Corpus than was seen in the Enron Corpus. Also, Naive Bayes outperforms its liblinear counterparts

| Method | Corpus | Feature Type | Web1T % | Score | Accuracy | Size(MB) |
|---|---|---|---|---|---|---|
| nb | enron | GM1 | 0 | 0.3998 | 0.6792 | 1.6986 |
| liblinear | enron | GM1 | 0 | 0.3388 | 0.8076 | 2.3840 |
| liblinear | enron | GM1 | 1 | 0.2528 | 0.8047 | 3.1834 |
| nb | enron | GM1 | 1 | 0.2030 | 0.6083 | 2.9969 |
| liblinear | enron | GM1 | 2 | 0.1424 | 0.8033 | 5.6421 |
| nb | enron | GM1 | 2 | 0.1113 | 0.6140 | 5.5165 |
| liblinear | enron | GM1 | 4 | 0.0776 | 0.8097 | 10.4301 |
| nb | enron | GM1 | 4 | 0.0593 | 0.6260 | 10.5528 |
| nb | enron | GM2 | 0 | 0.0437 | 0.7804 | 17.8759 |
| liblinear | enron | GM1 | 8 | 0.0397 | 0.8092 | 20.3672 |
| liblinear | enron | GM2 | 0 | 0.0383 | 0.8477 | 22.1173 |
| nb | enron | GM1 | 8 | 0.0300 | 0.6192 | 20.6262 |
| liblinear | enron | GM1 | 16 | 0.0203 | 0.8057 | 39.6953 |
| nb | enron | GM1 | 16 | 0.0154 | 0.6298 | 40.7717 |
| liblinear | enron | GM2 | 1 | 0.0142 | 0.8007 | 56.5585 |
| nb | enron | GB3 | 0 | 0.0131 | 0.7631 | 58.1904 |
| liblinear | enron | GB3 | 0 | 0.0119 | 0.8413 | 70.6541 |
| nb | enron | GM2 | 1 | 0.0106 | 0.6206 | 58.8138 |
| liblinear | enron | GB3 | 1 | 0.0083 | 0.8429 | 101.9446 |
| liblinear | enron | GM2 | 2 | 0.0072 | 0.8011 | 111.2368 |

Table 4.9: Highest Scoring Method-Feature Combinations Over All Groups for the Enron Email Corpus

in some situations. Just like with Enron, the highest scoring method-feature combination is not necessarily the most appropriate combination for deployment on a mobile phone. For the Twitter Corpus, liblinear using OSB3 and a Web1T%=0 has a .6127 accuracy with a size of 11.3466MB. This is the highest accuracy on the top 20 list that is still below 16MB. There are several accuracyies above 0.5 that have significantly smaller storage requirements.

## 4.5 Ability to Execute on an Android Mobile Phone

With scores calculated alongside accuracy and storage requirements, feasibility on a mobile device must be determined. The previous section clearly showed that $score = \frac{accuracy}{size}$ by itself does not provide an optimal solution for choosing an author detection method-feature combination on a mobile device. Tables 4.11 and 4.12 show the highest scoring method-feature

| Method | Corpus | Feature Type | Web1T % | Score | Accuracy | Size(MB) |
|---|---|---|---|---|---|---|
| nb | twitter | GM1 | 0 | 2.5176 | 0.5858 | 0.2327 |
| liblinear | twitter | GM1 | 0 | 1.5509 | 0.5806 | 0.3744 |
| nb | twitter | GM2 | 0 | 0.4482 | 0.5351 | 1.1939 |
| liblinear | twitter | GM2 | 0 | 0.2607 | 0.4524 | 1.7353 |
| liblinear | twitter | GM1 | 1 | 0.2244 | 0.5827 | 2.5972 |
| nb | twitter | GM1 | 1 | 0.1772 | 0.4565 | 2.5758 |
| nb | twitter | GB3 | 0 | 0.1729 | 0.5822 | 3.3669 |
| liblinear | twitter | GM1 | 2 | 0.1142 | 0.5773 | 5.0542 |
| nb | twitter | GM5 | 0 | 0.1006 | 0.3167 | 3.1482 |
| liblinear | twitter | GB3 | 0 | 0.0995 | 0.5074 | 5.0984 |
| nb | twitter | GM1 | 2 | 0.0927 | 0.4721 | 5.0936 |
| nb | twitter | OSB3 | 0 | 0.0797 | 0.6127 | 7.6833 |
| liblinear | twitter | GM5 | 0 | 0.0599 | 0.2488 | 4.1559 |
| liblinear | twitter | GM1 | 4 | 0.0581 | 0.5792 | 9.9645 |
| nb | twitter | GM1 | 4 | 0.0467 | 0.4728 | 10.1292 |
| liblinear | twitter | OSB3 | 0 | 0.0449 | 0.5100 | 11.3466 |
| liblinear | twitter | GM1 | 8 | 0.0296 | 0.5862 | 19.8001 |
| nb | twitter | GM1 | 8 | 0.0239 | 0.4823 | 20.2023 |
| liblinear | twitter | GM1 | 16 | 0.0147 | 0.5789 | 39.4154 |
| nb | twitter | GM1 | 16 | 0.0116 | 0.4694 | 40.3477 |

Table 4.10: Highest Scoring Method-Feature Combinations Over All Groups for the Twitter Short Message Corpus

combinations that have storage requirements less than 16MB and then ordered by accuracy. For the Enron Corpus Table 4.12 shows that the best accuracy achievable using the tools of thesis is 0.7735. For the Twitter Corpus, Table 4.12 shows that the best accuracy achievable using the tools of this thesis is 0.5525.

The Enron Email Corpus has 7 method-feature combinations with a storage requirement of less than 16MB. The Twitter Short Message Corpus has 14 method-feature combinations across all group sizes with a storage requirement under 16MB. Looking closely at the values of size and accuracy, there is little difference between the three highest accuracies in Table 4.11 but the third highest accuracy is more than double the size of the highest accuracy. That makes the choice of liblinear GM1 0 clearly the most appropriate choice for a mobile device. For the

| Method | Corpus | Feature Type | Web1T % | Score | Accuracy | Size(MB) | MLE | F-Score |
|---|---|---|---|---|---|---|---|---|
| liblinear | enron | GM1 | 0 | 0.1601 | 0.7735 | 4.83 | 0.3842 | 0.6257 |
| liblinear | enron | GM1 | 1 | 0.1113 | 0.7710 | 6.93 | 0.3859 | 0.6235 |
| liblinear | enron | GM1 | 2 | 0.0658 | 0.7704 | 11.71 | 0.3849 | 0.6255 |
| nb | enron | GM1 | 0 | 0.2954 | 0.6055 | 2.05 | 0.3842 | 0.3399 |
| nb | enron | GM1 | 4 | 0.0507 | 0.5640 | 11.12 | 0.3793 | 0.4771 |
| nb | enron | GM1 | 2 | 0.0908 | 0.5520 | 6.08 | 0.3849 | 0.4719 |
| nb | enron | GM1 | 1 | 0.1536 | 0.5462 | 3.56 | 0.3859 | 0.4649 |

Table 4.11: Highest Scoring Method-Feature Combinations Over All Groups for the Enron Email Corpus

| Method | Corpus | Feature Type | Web1T % | Score | Accuracy | Size(MB) | MLE | F-Score |
|---|---|---|---|---|---|---|---|---|
| nb | twitter | OSB3 | 0 | 0.0680 | 0.5525 | 8.13 | 0.1978 | 0.5338 |
| nb | twitter | GB3 | 0 | 0.1451 | 0.5203 | 3.59 | 0.1990 | 0.4820 |
| liblinear | twitter | GM1 | 1 | 0.0992 | 0.5159 | 5.20 | 0.1960 | 0.4953 |
| liblinear | twitter | GM1 | 0 | 0.6402 | 0.5141 | 0.80 | 0.1975 | 0.4934 |
| nb | twitter | GM1 | 0 | 1.8823 | 0.5140 | 0.27 | 0.1975 | 0.4708 |
| liblinear | twitter | GM1 | 2 | 0.0515 | 0.5131 | 9.97 | 0.1944 | 0.4915 |
| nb | twitter | GM2 | 0 | 0.3743 | 0.4750 | 1.27 | 0.1978 | 0.4350 |
| liblinear | twitter | GB3 | 0 | 0.0411 | 0.4522 | 11.01 | 0.1990 | 0.4207 |
| nb | twitter | GM1 | 4 | 0.0399 | 0.4070 | 10.20 | 0.1957 | 0.3838 |
| nb | twitter | GM1 | 2 | 0.0786 | 0.4059 | 5.16 | 0.1944 | 0.3833 |
| liblinear | twitter | GM2 | 0 | 0.1121 | 0.4002 | 3.57 | 0.1978 | 0.3726 |
| nb | twitter | GM1 | 1 | 0.1484 | 0.3922 | 2.64 | 0.1960 | 0.3672 |
| nb | twitter | GM5 | 0 | 0.0845 | 0.2726 | 3.22 | 0.1995 | 0.1902 |
| liblinear | twitter | GM5 | 0 | 0.0240 | 0.2095 | 8.73 | 0.1995 | 0.1547 |

Table 4.12: Highest Scoring Method-Feature Combinations Over All Groups for the Twitter Short Message Corpus

Twitter corpus, the top accuracy of .5525 for Naive Bayes using OSB3 is only slightly higher than 0.5203 for Naive Bayes using GB3 with a size that is which is less than half of OSB3. Naive Bayes using GB3 would be more appropriate for a mobile device.

The only remaining question is whether these method-feature combinations are stable performers across the group sizes. While standard deviation is one indicator, a plot of the accuracy, f-score, and MLE for each of these choice would be informative for consistent performance across group sizes. These plots can be compared to other method-feature combinations that have similiar accuracy and size values.

Figure 4.5 shows that liblinear GM1 has a steady decline from just above 0.8 to 0.6 from a groups size of 5 to a group size of 75. The accuracy for liblinear GM1 for the Enron corpus is virtually identical for groups sizes of 75 and 150. Figure 4.6 below shows that liblinear OSB3 for Twitter has declining accuracy from just above 0.6 to slightly above 0.2 as group size increases from 5 authors to 150 authors.

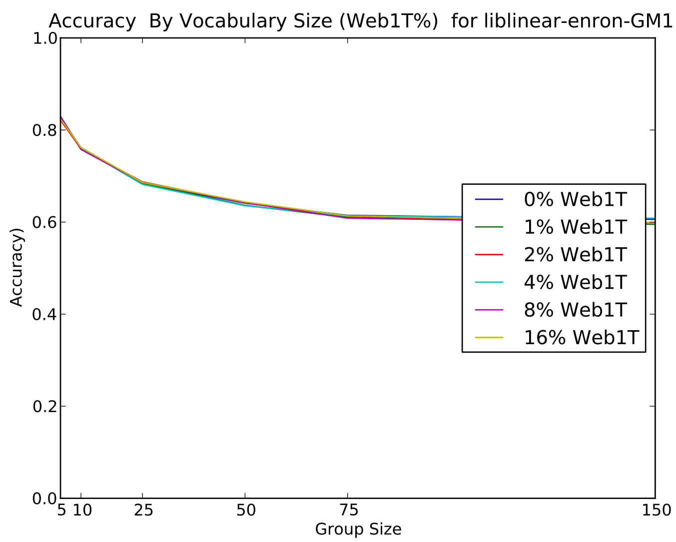



Figure 4.5: Accuracy Results over Group Size Using Liblinear GM1 for the Enron Email Corpus

The results for author detection over the Enron Email Corpus are far higher than for the Twitter Short Message Corpus for the selected method-feature combinations. This is not unexpected since results for the Enron Email Corpus have been higher than the Twitter Short Message Corpus across all test sets. With both selections having storage requirements of less than 1MB, execution of actual author detection on a mobile phone is practical as a next stage in future work.

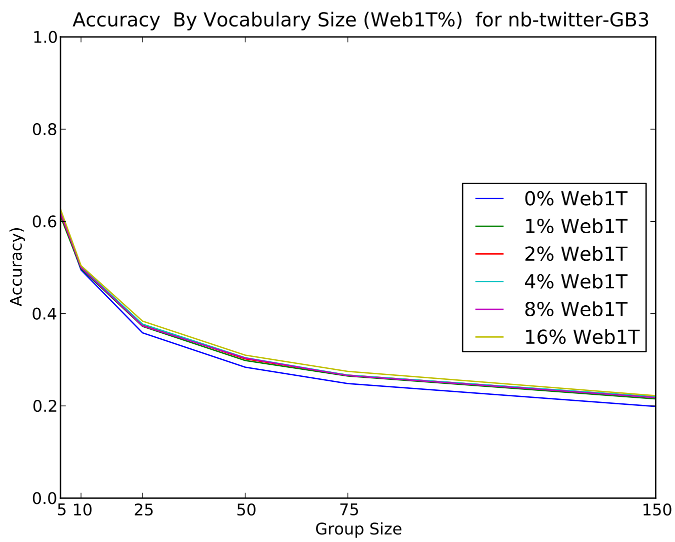



Figure 4.6: Accuracy Results over Group Size Using Liblinear OSB3 for the Twitter Short Message Corpus

THIS PAGE INTENTIONALLY LEFT BLANK

# CHAPTER 5:
# Conclusions and Future Work

## 5.1 Determine Accuracy and F-Score for Other Web1T Vocabulary Variations

This thesis only used model 0. Models such as 8 get rid of punctuation. Other models get rid of sentence boundaries, capitalization, etc. Reducing these allowed types reduces the total number of features, which could change accuracy and f-score.

## 5.2 Apply Good Turing or Witten-Bell Smoothing to Naive Bayes

Laplace smoothing using a +1 smoothing value for the bootstrap gave poor results. The Web1T smoothing was to resource intensive. Good Turing or Witten-Bell might improve Naive Bayes performance without the dramatic overhead of Web1T.

## 5.3 Increase the Twitter Short Message Size

The large difference in accuracy and F-Score between the ENRON Email Corpus and Twitter Short Message Corpus may be a function of how few tokens are present in the Twitter Corpus compared to the ENRON Email Corpus. If the most prolific Tweeters could be recorded for several months, a large enough body of tokens could be created to put some Tweeter's token count on par with the average ENRON email author's token count. That could clarify whether Twitter is inherently different from email or is simply less predictable when there is a smaller sample to analyze.

## 5.4 Rewrite LibLinear Data Structures

## 5.5 Placement on the Mobile Device

## 5.6 Study of how big a text/email social network usually is to appropriately size the to of the classifier group size

## 5.7 Study of Disk Storage to RAM Usage for Mobile Phones to support sizing for Dalvik VM or other OS RAM limitations

## 5.8 Statistical Study of Small-To-Large Versus Small-And-Large Groupings Results

Could take the most prolific speaker as measured in tokens and compare with most prolific speaker in terms of documents. Move each of those baseline speaker through different size groupings of other speakers.

## 5.9 Conduct LibLinear and Naive Bayes Tests Again with a Large ”Noise” Group

For Enron, could have 5 identfied authors and all the rest of the authors with a single author ID acting as the noise group. Cycle through all authors using the small-to-large and small-and-large grouping strategies. Repeat these experiments for group sizes of 10, 25, 50, and 75.

## 5.10 Actually Test the Top Scoring Method-Feature Combinations on Android Phones

# REFERENCES


[1] H. Love, *Attributing authorship: an introduction*. Cambridge University Press, Jun. 2002.

[2] E. Alpaydin, *Introduction to machine learning*. MIT Press, Oct. 2004.

[3] D. Jurafsky and J. H. Martin, *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall, 2009.

[4] V. Vapnik and C. Cortes, "Support-Vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995, 10.1023/A:1022627411411. [Online]. Available: http://dx.doi.org/10.1023/A:1022627411411

[5] R. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[6] "Multiclass SVMs," http://nlp.stanford.edu/IR-book/html/htmledition/multiclass-svms-1.html. [Online]. Available: http://nlp.stanford.edu/IR-book/html/htmledition/multiclass-svms-1.html

[7] T. Brants and A. Franz, "Web 1T 5-gram Version 1," 2006.

[8] D. Belazzougui, F. Botelho, and M. Dietzfelbinger, "Hash, displace, and compress," in *Algorithms - ESA 2009*, 2009, pp. 682–693. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-04128-0_61

[9] "CMPH - c minimal perfect hashing library," http://cmph.sourceforge.net/. [Online]. Available: http://cmph.sourceforge.net/

[10] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation," *AI 2006: Advances in Artificial Intelligence*, p. 10151021, 2006.

[11] "Gartner says worldwide mobile phone sales grew 17 per cent in first quarter 2010," http://www.gartner.com/it/page.jsp?id=1372013. [Online]. Available: http://www.gartner.com/it/page.jsp?id=1372013

[12] "BlackBerry - BlackBerry developer zone," http://us.blackberry.com/developers/. [Online]. Available: http://us.blackberry.com/developers/

[13] "Symbian SDKs," http://www.forum.nokia.com/info/sw.nokia.com/id/ec866fab-4b76-49f6-b5a5-af0631419e9c/S60_All_in_One_SDKs.html. [Online]. Available: http://www.forum.nokia.com/info/sw.nokia.com/id/ec866fab-4b76-49f6-b5a5-af0631419e9c/S60_All_in_One_SDKs.html

[14] M. L. Murphy, *Android Beyond Java*. CommonsWare, LLC, Sep. 2010.

[15] "Creating an iPhone application," http://developer.apple.com/library/ios/#referencelibrary/GettingStarted/Creating_an_iPhone_App/index.htm [Online]. Available: http://developer.apple.com/library/ios/#referencelibrary/GettingStarted/Creating_an_iPhone_App/index.html

[16] M. L. Murphy, *The Busy Coder's Guide to Android Development*. CommonsWare, Oct. 2010.

[17] "Enron email dataset," http://www-2.cs.cmu.edu/%7Eenron/. [Online]. Available: http://www-2.cs.cmu.edu/%7Eenron/

[18] "Streaming API documentation | dev.twitter.com," http://dev.twitter.com/pages/streaming_api. [Online]. Available: http://dev.twitter.com/pages/streaming_api

[19] D. M. Bikel and J. Sorensen, "If we want your opinion," in *Proceedings of the International Conference on Semantic Computing*. Washington, DC, USA: IEEE Computer Society, 2007, p. 493500, ACM ID: 1306375. [Online]. Available: http://portal.acm.org/citation.cfm?id=1304608.1306375

# APPENDIX A:
# LibLinear Accuracy and F-Score Results for the ENRON Email Corpus

| GM1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | 0.8269 | 0.9531 | 0.4815 | 0.0979 | 0.6864 | 0.9842 | 0.0000 | 0.2414 |
| | 1 | 0.8233 | 0.9578 | 0.4444 | 0.1003 | 0.6859 | 0.9826 | 0.0000 | 0.2384 |
| | 2 | 0.8216 | 0.9570 | 0.4444 | 0.0971 | 0.6881 | 0.9819 | 0.0000 | 0.2377 |
| | 4 | 0.8298 | 0.9590 | 0.4444 | 0.0949 | 0.6950 | 0.9821 | 0.0000 | 0.2315 |
| | 8 | 0.8298 | 0.9570 | 0.4444 | 0.0980 | 0.6878 | 0.9819 | 0.0000 | 0.2406 |
| | 16 | 0.8239 | 0.9732 | 0.4444 | 0.0987 | 0.6901 | 0.9878 | 0.0000 | 0.2316 |

Table A.1: liblinear-enron-GM1-ALL-ALL-5

| GM1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | 0.7611 | 0.9312 | 0.3776 | 0.1130 | 0.6122 | 0.9778 | 0.0000 | 0.2463 |
| | 1 | 0.7610 | 0.8890 | 0.3878 | 0.1109 | 0.6081 | 0.9699 | 0.0000 | 0.2490 |
| | 2 | 0.7594 | 0.9068 | 0.4388 | 0.1080 | 0.6093 | 0.9660 | 0.0000 | 0.2437 |
| | 4 | 0.7602 | 0.9086 | 0.4388 | 0.1074 | 0.6093 | 0.9692 | 0.0000 | 0.2451 |
| | 8 | 0.7578 | 0.9025 | 0.4388 | 0.1080 | 0.6113 | 0.9684 | 0.0000 | 0.2415 |
| | 16 | 0.7622 | 0.9187 | 0.3878 | 0.1142 | 0.6116 | 0.9698 | 0.0000 | 0.2425 |

Table A.2: liblinear-enron-GM1-ALL-ALL-10

| GM1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | 0.6845 | 0.8073 | 0.4430 | 0.1031 | 0.5251 | 0.9640 | 0.0000 | 0.2550 |
| | 1 | 0.6847 | 0.8064 | 0.4574 | 0.1071 | 0.5233 | 0.9572 | 0.0000 | 0.2567 |
| | 2 | 0.6873 | 0.8364 | 0.4500 | 0.1092 | 0.5256 | 0.9645 | 0.0000 | 0.2538 |
| | 4 | 0.6819 | 0.7836 | 0.4483 | 0.1057 | 0.5237 | 0.9558 | 0.0000 | 0.2554 |
| | 8 | 0.6862 | 0.8013 | 0.4599 | 0.1044 | 0.5291 | 0.9566 | 0.0000 | 0.2512 |
| | 16 | 0.6861 | 0.7925 | 0.4483 | 0.1033 | 0.5223 | 0.9568 | 0.0000 | 0.2620 |

Table A.3: liblinear-enron-GM1-ALL-ALL-25

| GM1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | 0.6411 | 0.7476 | 0.4341 | 0.0905 | 0.4800 | 0.9509 | 0.0000 | 0.2558 |
| | 1 | 0.6364 | 0.7280 | 0.4234 | 0.0982 | 0.4746 | 0.9561 | 0.0000 | 0.2571 |
| | 2 | 0.6420 | 0.7214 | 0.4287 | 0.0911 | 0.4764 | 0.9475 | 0.0000 | 0.2577 |
| | 4 | 0.6356 | 0.7052 | 0.4327 | 0.0888 | 0.4751 | 0.9532 | 0.0000 | 0.2578 |
| | 8 | 0.6419 | 0.7127 | 0.4376 | 0.0917 | 0.4780 | 0.9559 | 0.0000 | 0.2608 |
| | 16 | 0.6437 | 0.7524 | 0.4406 | 0.0913 | 0.4771 | 0.9504 | 0.0000 | 0.2576 |

Table A.4: liblinear-enron-GM1-ALL-ALL-50

| GM1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | 0.6146 | 0.7024 | 0.4155 | 0.0921 | 0.4492 | 0.9437 | 0.0000 | 0.2588 |
| | 1 | 0.6101 | 0.7011 | 0.3880 | 0.1031 | 0.4407 | 0.9453 | 0.0000 | 0.2633 |
| | 2 | 0.6127 | 0.6858 | 0.3995 | 0.0978 | 0.4462 | 0.9511 | 0.0000 | 0.2584 |
| | 4 | 0.6132 | 0.6814 | 0.4006 | 0.0969 | 0.4417 | 0.9402 | 0.0000 | 0.2609 |
| | 8 | 0.6085 | 0.6836 | 0.4074 | 0.0921 | 0.4432 | 0.9392 | 0.0000 | 0.2574 |
| | 16 | 0.6137 | 0.6716 | 0.4030 | 0.0948 | 0.4403 | 0.9413 | 0.0000 | 0.2595 |

Table A.5: liblinear-enron-GM1-ALL-ALL-75

| GM1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | 0.6060 | 0.6074 | 0.6033 | 0.0020 | 0.4000 | 0.9316 | 0.0000 | 0.2610 |
| | 1 | 0.5951 | 0.6155 | 0.5849 | 0.0144 | 0.3968 | 0.9402 | 0.0000 | 0.2678 |
| | 2 | 0.5982 | 0.6049 | 0.5949 | 0.0047 | 0.3958 | 0.9389 | 0.0000 | 0.2676 |
| | 4 | 0.6083 | 0.6093 | 0.6065 | 0.0013 | 0.4037 | 0.9488 | 0.0000 | 0.2640 |
| | 8 | 0.5990 | 0.6008 | 0.5982 | 0.0012 | 0.4023 | 0.9451 | 0.0000 | 0.2639 |
| | 16 | 0.5987 | 0.6011 | 0.5975 | 0.0017 | 0.3991 | 0.9489 | 0.0000 | 0.2664 |

Table A.6: liblinear-enron-GM1-ALL-ALL-150

| GM2 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | 0.8607 | 0.9753 | 0.5185 | 0.0980 | 0.7309 | 1.0000 | 0.0000 | 0.2402 |
| | 1 | 0.8193 | 0.9544 | 0.4444 | 0.1034 | 0.6761 | 0.9781 | 0.0000 | 0.2389 |
| | 2 | 0.8192 | 0.9448 | 0.4444 | 0.1004 | 0.6782 | 0.9778 | 0.0000 | 0.2400 |
| | 4 | 0.8187 | 0.9560 | 0.4444 | 0.1014 | 0.6747 | 0.9834 | 0.0000 | 0.2412 |
| | 8 | 0.8199 | 0.9547 | 0.4444 | 0.1024 | 0.6747 | 0.9817 | 0.0000 | 0.2419 |
| | 16 | 0.8154 | 0.9606 | 0.4444 | 0.1037 | 0.6782 | 0.9881 | 0.0000 | 0.2379 |

Table A.7: liblinear-enron-GM2-ALL-ALL-5

| GM2 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | 0.8150 | 0.9369 | 0.5000 | 0.1044 | 0.6869 | 0.9811 | 0.0000 | 0.2443 |
| | 1 | 0.7551 | 0.9093 | 0.3936 | 0.1114 | 0.5942 | 0.9711 | 0.0000 | 0.2510 |
| | 2 | 0.7578 | 0.9255 | 0.3936 | 0.1129 | 0.6001 | 0.9711 | 0.0000 | 0.2510 |
| | 4 | 0.7502 | 0.8969 | 0.3936 | 0.1126 | 0.5998 | 0.9675 | 0.0000 | 0.2480 |
| | 8 | 0.7581 | 0.8739 | 0.3936 | 0.1122 | 0.6003 | 0.9678 | 0.0000 | 0.2488 |
| | 16 | 0.7528 | 0.9065 | 0.3936 | 0.1138 | 0.5986 | 0.9728 | 0.0000 | 0.2510 |

Table A.8: liblinear-enron-GM2-ALL-ALL-10

| GM2 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | 0.7696 | 0.8989 | 0.4824 | 0.1111 | 0.6317 | 0.9814 | 0.0000 | 0.2630 |
| | 1 | 0.6790 | 0.8075 | 0.4512 | 0.1045 | 0.5171 | 0.9632 | 0.0000 | 0.2591 |
| | 2 | 0.6822 | 0.8065 | 0.4562 | 0.1023 | 0.5212 | 0.9612 | 0.0000 | 0.2558 |
| | 4 | 0.6810 | 0.7994 | 0.4465 | 0.1055 | 0.5174 | 0.9674 | 0.0000 | 0.2565 |

Table A.9: liblinear-enron-GM2-ALL-ALL-25

| GM2 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | 0.7402 | 0.8240 | 0.5252 | 0.0987 | 0.6002 | 0.9780 | 0.0000 | 0.2678 |
| | 1 | 0.6329 | 0.7216 | 0.4211 | 0.0925 | 0.4682 | 0.9505 | 0.0000 | 0.2616 |
| | 2 | 0.6341 | 0.7156 | 0.4119 | 0.0963 | 0.4709 | 0.9504 | 0.0000 | 0.2581 |

Table A.10: liblinear-enron-GM2-ALL-ALL-50

| GM2 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | 0.7330 | 0.7969 | 0.5437 | 0.0867 | 0.5832 | 0.9786 | 0.0000 | 0.2721 |
| | 1 | 0.6120 | 0.6959 | 0.3872 | 0.1035 | 0.4454 | 0.9564 | 0.0000 | 0.2650 |
| | 2 | 0.5987 | 0.6677 | 0.3874 | 0.0974 | 0.4385 | 0.9523 | 0.0000 | 0.2653 |

Table A.11: liblinear-enron-GM2-ALL-ALL-75

| GM2 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | 0.7447 | 0.7456 | 0.7429 | 0.0013 | 0.5516 | 0.9737 | 0.0000 | 0.2791 |
| | 1 | 0.5978 | 0.6047 | 0.5841 | 0.0097 | 0.3979 | 0.9387 | 0.0000 | 0.2697 |

Table A.12: liblinear-enron-GM2-ALL-ALL-150

| GM5 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | 0.6881 | 0.9636 | 0.3017 | 0.1576 | 0.5062 | 1.0000 | 0.0000 | 0.3012 |
| | 1 | 0.8117 | 0.9685 | 0.4000 | 0.1167 | 0.6773 | 0.9869 | 0.0000 | 0.2407 |
| | 2 | 0.8118 | 0.9550 | 0.4000 | 0.1133 | 0.6725 | 0.9836 | 0.0000 | 0.2423 |
| | 4 | 0.8130 | 0.9455 | 0.4000 | 0.1142 | 0.6772 | 0.9821 | 0.0000 | 0.2398 |
| | 8 | 0.8070 | 0.9513 | 0.4000 | 0.1152 | 0.6749 | 0.9824 | 0.0000 | 0.2367 |

Table A.13: liblinear-enron-GM5-ALL-ALL-5

| GM5 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | 0.6297 | 0.8560 | 0.2548 | 0.1432 | 0.4605 | 0.9870 | 0.0000 | 0.2997 |
| | 1 | 0.7519 | 0.9022 | 0.3908 | 0.1141 | 0.5979 | 0.9782 | 0.0000 | 0.2478 |
| | 2 | 0.7440 | 0.9221 | 0.3908 | 0.1128 | 0.5981 | 0.9733 | 0.0000 | 0.2487 |
| | 4 | 0.7418 | 0.9256 | 0.3908 | 0.1161 | 0.5992 | 0.9729 | 0.0000 | 0.2479 |

Table A.14: liblinear-enron-GM5-ALL-ALL-10

| GM5 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | 0.5499 | 0.7076 | 0.3305 | 0.1318 | 0.4372 | 1.0000 | 0.0000 | 0.3073 |
| | 1 | 0.6759 | 0.7867 | 0.4371 | 0.1075 | 0.5242 | 0.9554 | 0.0000 | 0.2557 |
| | 2 | 0.6728 | 0.7759 | 0.4371 | 0.0982 | 0.5252 | 0.9599 | 0.0000 | 0.2552 |

Table A.15: liblinear-enron-GM5-ALL-ALL-25

| GM5 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | 0.5323 | 0.6564 | 0.3088 | 0.1049 | 0.4262 | 0.9870 | 0.0000 | 0.3097 |
| | 1 | 0.6259 | 0.7342 | 0.4304 | 0.0929 | 0.4757 | 0.9572 | 0.0000 | 0.2655 |

Table A.16: liblinear-enron-GM5-ALL-ALL-50

| GM5 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | 0.5211 | 0.6206 | 0.3240 | 0.0953 | 0.4148 | 0.9870 | 0.0000 | 0.3171 |

Table A.17: liblinear-enron-GM5-ALL-ALL-75

| GB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | 0.8529 | 0.9835 | 0.5185 | 0.1014 | 0.7203 | 1.0000 | 0.0000 | 0.2469 |
| | 1 | 0.8494 | 0.9673 | 0.5185 | 0.1016 | 0.7172 | 0.9854 | 0.0000 | 0.2459 |
| | 2 | 0.8476 | 0.9805 | 0.5185 | 0.1040 | 0.7152 | 0.9890 | 0.0000 | 0.2470 |
| | 4 | 0.8579 | 0.9762 | 0.5185 | 0.1007 | 0.7174 | 0.9844 | 0.0000 | 0.2516 |
| | 8 | 0.8536 | 0.9786 | 0.5185 | 0.1003 | 0.7152 | 0.9921 | 0.0000 | 0.2501 |
| | 16 | 0.8523 | 0.9756 | 0.5185 | 0.1028 | 0.7136 | 0.9886 | 0.0000 | 0.2520 |

Table A.18: liblinear-enron-GB3-ALL-ALL-5

| GB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | 0.8124 | 0.9538 | 0.4787 | 0.1149 | 0.6699 | 1.0000 | 0.0000 | 0.2599 |
| | 1 | 0.8127 | 0.9341 | 0.4894 | 0.1084 | 0.6712 | 1.0000 | 0.0000 | 0.2545 |
| | 2 | 0.8128 | 0.9297 | 0.4894 | 0.1074 | 0.6753 | 0.9870 | 0.0000 | 0.2509 |
| | 4 | 0.8096 | 0.9426 | 0.4894 | 0.1079 | 0.6723 | 1.0000 | 0.0000 | 0.2548 |
| | 8 | 0.8134 | 0.9512 | 0.4894 | 0.1100 | 0.6725 | 0.9870 | 0.0000 | 0.2545 |

Table A.19: liblinear-enron-GB3-ALL-ALL-10

| GB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | 0.7680 | 0.8882 | 0.5215 | 0.1068 | 0.6218 | 0.9797 | 0.0000 | 0.2675 |
| | 1 | 0.7652 | 0.8744 | 0.5158 | 0.1136 | 0.6188 | 0.9816 | 0.0000 | 0.2657 |
| | 2 | 0.7684 | 0.8788 | 0.5130 | 0.1119 | 0.6208 | 0.9772 | 0.0000 | 0.2646 |

Table A.20: liblinear-enron-GB3-ALL-ALL-25

| GB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | 0.7409 | 0.8465 | 0.4980 | 0.1063 | 0.5914 | 1.0000 | 0.0000 | 0.2725 |
| | 1 | 0.7372 | 0.8204 | 0.4731 | 0.1083 | 0.5865 | 0.9753 | 0.0000 | 0.2732 |

Table A.21: liblinear-enron-GB3-ALL-ALL-50

| GB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | 0.7300 | 0.7955 | 0.5220 | 0.0947 | 0.5710 | 0.9870 | 0.0000 | 0.2783 |
| | 1 | 0.7336 | 0.8161 | 0.5273 | 0.0952 | 0.5773 | 0.9763 | 0.0000 | 0.2722 |

Table A.22: liblinear-enron-GB3-ALL-ALL-75

| OSB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | 0.8690 | 0.9732 | 0.5185 | 0.0928 | 0.7386 | 1.0000 | 0.0000 | 0.2435 |
| | 1 | 0.8667 | 0.9741 | 0.5185 | 0.0964 | 0.7375 | 0.9921 | 0.0000 | 0.2396 |
| | 2 | 0.8645 | 0.9765 | 0.5185 | 0.0991 | 0.7369 | 0.9923 | 0.0000 | 0.2424 |
| | 4 | 0.8687 | 0.9762 | 0.5185 | 0.0958 | 0.7367 | 0.9884 | 0.0000 | 0.2416 |

Table A.23: liblinear-enron-OSB3-ALL-ALL-5

| OSB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | 0.8250 | 0.9469 | 0.5106 | 0.1028 | 0.6886 | 0.9867 | 0.0000 | 0.2502 |

Table A.24: liblinear-enron-OSB3-ALL-ALL-10

| OSB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | 0.7826 | 0.8954 | 0.5385 | 0.1038 | 0.6374 | 0.9815 | 0.0000 | 0.2599 |

Table A.25: liblinear-enron-OSB3-ALL-ALL-25

| OSB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | 0.7567 | 0.8569 | 0.5160 | 0.1016 | 0.6056 | 0.9793 | 0.0000 | 0.2645 |

Table A.26: liblinear-enron-OSB3-ALL-ALL-50

| OSB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | 0.7470 | 0.7931 | 0.5547 | 0.0862 | 0.5858 | 0.9786 | 0.0000 | 0.2703 |

Table A.27: liblinear-enron-OSB3-ALL-ALL-75

# APPENDIX B:
# LibLinear Accuracy and F-Score Results for the Twitter Short Message Corpus

| GM1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | 0.6212 | 0.8089 | 0.4737 | 0.0739 | 0.6023 | 0.9696 | 0.1791 | 0.1416 |
| | 1 | 0.6228 | 0.8211 | 0.4713 | 0.0778 | 0.6034 | 0.9597 | 0.1429 | 0.1445 |
| | 2 | 0.6147 | 0.8966 | 0.4218 | 0.0888 | 0.5948 | 0.9697 | 0.0000 | 0.1514 |
| | 4 | 0.6172 | 0.8546 | 0.3846 | 0.0850 | 0.5992 | 0.9811 | 0.1200 | 0.1450 |
| | 8 | 0.6273 | 0.9026 | 0.4661 | 0.0813 | 0.6087 | 0.9735 | 0.1404 | 0.1419 |
| | 16 | 0.6181 | 0.8324 | 0.4458 | 0.0816 | 0.6013 | 0.9600 | 0.1515 | 0.1386 |

Table B.1: liblinear-twitter-GM1-ALL-ALL-5

| GM1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | 0.4762 | 0.6234 | 0.3448 | 0.0725 | 0.4537 | 0.9556 | 0.0385 | 0.1607 |
| | 1 | 0.4813 | 0.6389 | 0.3627 | 0.0662 | 0.4605 | 0.9482 | 0.0755 | 0.1592 |
| | 2 | 0.4845 | 0.6567 | 0.3358 | 0.0699 | 0.4617 | 0.9699 | 0.0702 | 0.1615 |
| | 4 | 0.4841 | 0.6900 | 0.3184 | 0.0750 | 0.4628 | 0.9517 | 0.0299 | 0.1639 |
| | 8 | 0.4816 | 0.7194 | 0.3080 | 0.0784 | 0.4576 | 0.9621 | 0.0000 | 0.1637 |
| | 16 | 0.4800 | 0.6362 | 0.2846 | 0.0723 | 0.4567 | 0.9474 | 0.0370 | 0.1613 |

Table B.2: liblinear-twitter-GM1-ALL-ALL-10

| GM1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | 0.3461 | 0.4816 | 0.2735 | 0.0562 | 0.3197 | 0.9344 | 0.0000 | 0.1697 |
| | 1 | 0.3408 | 0.4390 | 0.2408 | 0.0596 | 0.3160 | 0.9221 | 0.0000 | 0.1740 |
| | 2 | 0.3465 | 0.4309 | 0.2714 | 0.0430 | 0.3225 | 0.9225 | 0.0000 | 0.1717 |
| | 4 | 0.3510 | 0.4402 | 0.2811 | 0.0476 | 0.3264 | 0.9358 | 0.0000 | 0.1724 |
| | 8 | 0.3419 | 0.4296 | 0.2591 | 0.0483 | 0.3189 | 0.9231 | 0.0000 | 0.1707 |
| | 16 | 0.3411 | 0.4296 | 0.2651 | 0.0417 | 0.3133 | 0.9011 | 0.0000 | 0.1705 |

Table B.3: liblinear-twitter-GM1-ALL-ALL-25

| GM1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | 0.2693 | 0.3371 | 0.2190 | 0.0369 | 0.2496 | 0.8798 | 0.0000 | 0.1754 |
| | 1 | 0.2704 | 0.3173 | 0.2219 | 0.0365 | 0.2483 | 0.9153 | 0.0000 | 0.1747 |
| | 2 | 0.2705 | 0.3338 | 0.2180 | 0.0415 | 0.2419 | 0.8922 | 0.0000 | 0.1753 |
| | 4 | 0.2710 | 0.3272 | 0.2032 | 0.0457 | 0.2464 | 0.8889 | 0.0000 | 0.1716 |
| | 8 | 0.2712 | 0.3171 | 0.2274 | 0.0281 | 0.2431 | 0.9119 | 0.0000 | 0.1727 |
| | 16 | 0.2710 | 0.3326 | 0.2344 | 0.0308 | 0.2458 | 0.9035 | 0.0000 | 0.1713 |

Table B.4: liblinear-twitter-GM1-ALL-ALL-50

| GM1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | 0.2273 | 0.2738 | 0.1810 | 0.0371 | 0.2100 | 0.8687 | 0.0000 | 0.1713 |
| | 1 | 0.2293 | 0.2626 | 0.1768 | 0.0320 | 0.2092 | 0.9157 | 0.0000 | 0.1718 |
| | 2 | 0.2318 | 0.2730 | 0.1818 | 0.0364 | 0.2127 | 0.8750 | 0.0000 | 0.1729 |
| | 4 | 0.2310 | 0.2776 | 0.1874 | 0.0313 | 0.2101 | 0.8971 | 0.0000 | 0.1687 |
| | 8 | 0.2354 | 0.2910 | 0.1743 | 0.0406 | 0.2097 | 0.8873 | 0.0000 | 0.1720 |
| | 16 | 0.2368 | 0.2780 | 0.1910 | 0.0361 | 0.2091 | 0.9037 | 0.0000 | 0.1740 |

Table B.5: liblinear-twitter-GM1-ALL-ALL-75

| GM1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | 0.1851 | 0.1875 | 0.1802 | 0.0034 | 0.1629 | 0.8582 | 0.0000 | 0.1623 |
| | 1 | 0.1888 | 0.1932 | 0.1802 | 0.0061 | 0.1642 | 0.8212 | 0.0000 | 0.1636 |
| | 2 | 0.1829 | 0.1833 | 0.1821 | 0.0006 | 0.1595 | 0.7807 | 0.0000 | 0.1575 |
| | 4 | 0.1921 | 0.1943 | 0.1910 | 0.0016 | 0.1665 | 0.8792 | 0.0000 | 0.1594 |
| | 8 | 0.1893 | 0.1913 | 0.1884 | 0.0014 | 0.1640 | 0.8139 | 0.0000 | 0.1650 |
| | 16 | 0.1877 | 0.1883 | 0.1865 | 0.0008 | 0.1644 | 0.8143 | 0.0000 | 0.1614 |

Table B.6: liblinear-twitter-GM1-ALL-ALL-150

| GM2 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | 0.4844 | 0.7282 | 0.2664 | 0.0887 | 0.4501 | 0.9272 | 0.0000 | 0.1676 |
| | 1 | 0.6241 | 0.8454 | 0.4664 | 0.0780 | 0.6029 | 0.9886 | 0.1159 | 0.1482 |
| | 2 | 0.6221 | 0.8232 | 0.3934 | 0.0867 | 0.6054 | 0.9697 | 0.1071 | 0.1423 |
| | 4 | 0.6253 | 0.8210 | 0.4245 | 0.0791 | 0.6077 | 0.9545 | 0.1639 | 0.1405 |
| | 8 | 0.6282 | 0.8489 | 0.4773 | 0.0734 | 0.6130 | 0.9773 | 0.1818 | 0.1373 |
| | 16 | 0.6258 | 0.8544 | 0.4094 | 0.0877 | 0.6119 | 0.9603 | 0.2157 | 0.1348 |

Table B.7: liblinear-twitter-GM2-ALL-ALL-5

| GM2 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | 0.3700 | 0.5725 | 0.1985 | 0.0775 | 0.3455 | 0.9150 | 0.0000 | 0.1686 |
| | 1 | 0.4904 | 0.6219 | 0.3429 | 0.0665 | 0.4657 | 0.9575 | 0.0000 | 0.1636 |
| | 2 | 0.4903 | 0.6491 | 0.3560 | 0.0628 | 0.4689 | 0.9549 | 0.0879 | 0.1585 |
| | 4 | 0.4962 | 0.6924 | 0.3379 | 0.0711 | 0.4711 | 0.9771 | 0.0000 | 0.1639 |
| | 8 | 0.4842 | 0.6693 | 0.3593 | 0.0623 | 0.4622 | 0.9524 | 0.0857 | 0.1603 |
| | 16 | 0.4891 | 0.6961 | 0.3142 | 0.0737 | 0.4684 | 0.9421 | 0.0000 | 0.1581 |

Table B.8: liblinear-twitter-GM2-ALL-ALL-10

| GM2 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | 0.2641 | 0.3285 | 0.1882 | 0.0346 | 0.2437 | 0.8473 | 0.0000 | 0.1734 |
| | 1 | 0.3468 | 0.4945 | 0.2702 | 0.0518 | 0.3226 | 0.9299 | 0.0000 | 0.1737 |
| | 2 | 0.3464 | 0.4433 | 0.2642 | 0.0508 | 0.3219 | 0.9542 | 0.0000 | 0.1678 |
| | 4 | 0.3526 | 0.4450 | 0.2665 | 0.0470 | 0.3304 | 0.9438 | 0.0000 | 0.1692 |

Table B.9: liblinear-twitter-GM2-ALL-ALL-25

| GM2 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | 0.2097 | 0.2745 | 0.1587 | 0.0304 | 0.2006 | 0.8333 | 0.0000 | 0.1732 |
| | 1 | 0.2679 | 0.3381 | 0.2275 | 0.0404 | 0.2453 | 0.8750 | 0.0000 | 0.1712 |
| | 2 | 0.2707 | 0.3221 | 0.2116 | 0.0373 | 0.2470 | 0.8777 | 0.0000 | 0.1796 |

Table B.10: liblinear-twitter-GM2-ALL-ALL-50

| GM2 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | 0.1800 | 0.2186 | 0.1277 | 0.0302 | 0.1729 | 0.7926 | 0.0000 | 0.1664 |
| | 1 | 0.2350 | 0.2861 | 0.1960 | 0.0331 | 0.2147 | 0.8914 | 0.0000 | 0.1715 |
| | 2 | 0.2354 | 0.2697 | 0.1867 | 0.0293 | 0.2130 | 0.8397 | 0.0000 | 0.1710 |

Table B.11: liblinear-twitter-GM2-ALL-ALL-75

| GM2 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | 0.1525 | 0.1554 | 0.1466 | 0.0042 | 0.1439 | 0.8092 | 0.0000 | 0.1605 |
| | 1 | 0.1889 | 0.1893 | 0.1882 | 0.0006 | 0.1659 | 0.7742 | 0.0000 | 0.1627 |

Table B.12: liblinear-twitter-GM2-ALL-ALL-150

| GM5 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | 0.2764 | 0.5119 | 0.1687 | 0.0560 | 0.1995 | 0.6715 | 0.0000 | 0.1509 |
| | 1 | 0.5868 | 0.8362 | 0.4040 | 0.0830 | 0.5657 | 0.9498 | 0.1667 | 0.1468 |
| | 2 | 0.5802 | 0.7639 | 0.3689 | 0.0835 | 0.5594 | 0.9457 | 0.0845 | 0.1488 |

Table B.13: liblinear-twitter-GM5-ALL-ALL-5

| GM5 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | 0.1718 | 0.2811 | 0.0961 | 0.0371 | 0.1238 | 0.5625 | 0.0000 | 0.1172 |
| | 1 | 0.4383 | 0.5825 | 0.3247 | 0.0648 | 0.4150 | 0.9237 | 0.0519 | 0.1609 |
| | 2 | 0.4382 | 0.5657 | 0.3108 | 0.0643 | 0.4182 | 0.9302 | 0.0000 | 0.1577 |

Table B.14: liblinear-twitter-GM5-ALL-ALL-10

| GM5 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | 0.1060 | 0.1677 | 0.0717 | 0.0254 | 0.0886 | 0.6154 | 0.0000 | 0.1141 |
| | 1 | 0.3011 | 0.3718 | 0.2378 | 0.0356 | 0.2784 | 0.9105 | 0.0000 | 0.1619 |

Table B.15: liblinear-twitter-GM5-ALL-ALL-25

| GM5 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | 0.0805 | 0.1060 | 0.0666 | 0.0149 | 0.0792 | 0.5373 | 0.0000 | 0.1124 |

Table B.16: liblinear-twitter-GM5-ALL-ALL-50

| GM5 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | 0.0643 | 0.0792 | 0.0421 | 0.0123 | 0.0658 | 0.5392 | 0.0000 | 0.1018 |

Table B.17: liblinear-twitter-GM5-ALL-ALL-75

| GM5 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | 0.0636 | 0.0667 | 0.0573 | 0.0044 | 0.0707 | 0.5443 | 0.0000 | 0.1076 |

Table B.18: liblinear-twitter-GM5-ALL-ALL-150

| GB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | 0.5405 | 0.8313 | 0.3737 | 0.0845 | 0.5059 | 0.9360 | 0.0000 | 0.1665 |
| | 1 | 0.5312 | 0.7360 | 0.3734 | 0.0752 | 0.4996 | 0.9308 | 0.0000 | 0.1684 |
| | 2 | 0.5447 | 0.7269 | 0.3850 | 0.0702 | 0.5125 | 0.9375 | 0.0000 | 0.1630 |
| | 4 | 0.5399 | 0.7538 | 0.3571 | 0.0834 | 0.5093 | 0.9354 | 0.0000 | 0.1684 |
| | 8 | 0.5404 | 0.8297 | 0.3908 | 0.0842 | 0.5085 | 0.9290 | 0.0000 | 0.1665 |
| | 16 | 0.5386 | 0.7564 | 0.3780 | 0.0768 | 0.5020 | 0.9416 | 0.0000 | 0.1740 |

Table B.19: liblinear-twitter-GB3-ALL-ALL-5

| GB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | 0.4231 | 0.5716 | 0.3021 | 0.0666 | 0.3913 | 0.9302 | 0.0000 | 0.1732 |
| | 1 | 0.4207 | 0.5795 | 0.2998 | 0.0632 | 0.3868 | 0.8806 | 0.0000 | 0.1734 |
| | 2 | 0.4221 | 0.6907 | 0.3259 | 0.0717 | 0.3938 | 0.9049 | 0.0000 | 0.1711 |
| | 4 | 0.4226 | 0.5960 | 0.3114 | 0.0682 | 0.3899 | 0.9231 | 0.0000 | 0.1757 |
| | 8 | 0.4246 | 0.5868 | 0.3045 | 0.0706 | 0.3961 | 0.9266 | 0.0000 | 0.1732 |

Table B.20: liblinear-twitter-GB3-ALL-ALL-10

| GB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | 0.3123 | 0.4116 | 0.2218 | 0.0525 | 0.2865 | 0.8750 | 0.0000 | 0.1770 |
| | 1 | 0.3087 | 0.4029 | 0.2379 | 0.0469 | 0.2837 | 0.9147 | 0.0000 | 0.1726 |
| | 2 | 0.3134 | 0.4094 | 0.2569 | 0.0423 | 0.2829 | 0.8949 | 0.0000 | 0.1782 |

Table B.21: liblinear-twitter-GB3-ALL-ALL-25

| GB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | 0.2467 | 0.2968 | 0.1960 | 0.0322 | 0.2240 | 0.8992 | 0.0000 | 0.1676 |
| | 1 | 0.2465 | 0.3082 | 0.1949 | 0.0353 | 0.2250 | 0.8864 | 0.0000 | 0.1729 |

Table B.22: liblinear-twitter-GB3-ALL-ALL-50

| GB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | 0.2132 | 0.2497 | 0.1842 | 0.0224 | 0.1963 | 0.8530 | 0.0000 | 0.1675 |
| | 1 | 0.2155 | 0.2477 | 0.1803 | 0.0310 | 0.1916 | 0.8803 | 0.0000 | 0.1659 |

Table B.23: liblinear-twitter-GB3-ALL-ALL-75

| GB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | 0.1739 | 0.1825 | 0.1696 | 0.0061 | 0.1514 | 0.8357 | 0.0000 | 0.1507 |

Table B.24: liblinear-twitter-GB3-ALL-ALL-150

| OSB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | 0.5430 | 0.7559 | 0.3680 | 0.0818 | 0.5144 | 0.9513 | 0.0000 | 0.1641 |
| | 1 | 0.5391 | 0.7747 | 0.4000 | 0.0729 | 0.5062 | 0.9362 | 0.0000 | 0.1647 |
| | 2 | 0.5427 | 0.7651 | 0.3731 | 0.0790 | 0.5084 | 0.9425 | 0.0000 | 0.1687 |
| | 4 | 0.5391 | 0.7747 | 0.3934 | 0.0793 | 0.5085 | 0.9434 | 0.0000 | 0.1673 |

Table B.25: liblinear-twitter-OSB3-ALL-ALL-5

| OSB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | 0.4271 | 0.5520 | 0.3216 | 0.0587 | 0.3987 | 0.9261 | 0.0000 | 0.1669 |
| | 1 | 0.4288 | 0.5847 | 0.3219 | 0.0646 | 0.3973 | 0.9453 | 0.0000 | 0.1774 |
| | 2 | 0.4255 | 0.5802 | 0.3280 | 0.0579 | 0.3936 | 0.9125 | 0.0000 | 0.1768 |

Table B.26: liblinear-twitter-OSB3-ALL-ALL-10

| OSB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | 0.3084 | 0.4086 | 0.2331 | 0.0532 | 0.2849 | 0.8731 | 0.0000 | 0.1766 |

Table B.27: liblinear-twitter-OSB3-ALL-ALL-25

| OSB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | 0.2520 | 0.2913 | 0.2023 | 0.0296 | 0.2293 | 0.8686 | 0.0000 | 0.1728 |

Table B.28: liblinear-twitter-OSB3-ALL-ALL-50

| OSB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | 0.2211 | 0.2493 | 0.1815 | 0.0254 | 0.2009 | 0.8839 | 0.0000 | 0.1709 |

Table B.29: liblinear-twitter-OSB3-ALL-ALL-75

| OSB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | 0.1750 | 0.1839 | 0.1705 | 0.0063 | 0.1580 | 0.8239 | 0.0000 | 0.1531 |

Table B.30: liblinear-twitter-OSB3-ALL-ALL-150

THIS PAGE INTENTIONALLY LEFT BLANK

# APPENDIX C:
# Naive Bayes Accuracy and F-Score Results for the ENRON Email Corpus

| GM1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | 0.7215 | 0.9114 | 0.4815 | 0.0960 | 0.4350 | 0.9730 | 0.0000 | 0.3637 |
| | 1 | 0.6441 | 0.8864 | 0.2937 | 0.1462 | 0.5404 | 0.9453 | 0.0000 | 0.2494 |
| | 2 | 0.6505 | 0.8877 | 0.2256 | 0.1460 | 0.5510 | 0.9467 | 0.0000 | 0.2474 |
| | 4 | 0.6610 | 0.8724 | 0.2898 | 0.1378 | 0.5526 | 0.9483 | 0.0000 | 0.2501 |
| | 8 | 0.6534 | 0.8864 | 0.2950 | 0.1387 | 0.5461 | 0.9494 | 0.0000 | 0.2482 |
| | 16 | 0.6663 | 0.8698 | 0.2551 | 0.1438 | 0.5567 | 0.9513 | 0.0000 | 0.2472 |

Table C.1: nb-enron-GM1-ALL-ALL-5

| GM1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | 0.5768 | 0.7663 | 0.3311 | 0.1086 | 0.3121 | 0.9164 | 0.0000 | 0.3137 |
| | 1 | 0.5189 | 0.7117 | 0.2923 | 0.1220 | 0.4421 | 0.9655 | 0.0000 | 0.2343 |
| | 2 | 0.5215 | 0.7192 | 0.2904 | 0.1215 | 0.4446 | 0.9157 | 0.0000 | 0.2344 |
| | 4 | 0.5406 | 0.7545 | 0.2715 | 0.1269 | 0.4554 | 0.9500 | 0.0000 | 0.2372 |
| | 8 | 0.5349 | 0.7164 | 0.2647 | 0.1192 | 0.4534 | 0.9157 | 0.0000 | 0.2351 |
| | 16 | 0.5377 | 0.7174 | 0.2763 | 0.1209 | 0.4537 | 0.9500 | 0.0000 | 0.2354 |

Table C.2: nb-enron-GM1-ALL-ALL-10

| GM1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | 0.4083 | 0.5192 | 0.2996 | 0.0581 | 0.1852 | 0.8796 | 0.0000 | 0.2424 |
| | 1 | 0.3956 | 0.5915 | 0.2745 | 0.0983 | 0.3457 | 0.9870 | 0.0000 | 0.2139 |
| | 2 | 0.4037 | 0.5964 | 0.2822 | 0.0958 | 0.3488 | 0.9870 | 0.0000 | 0.2137 |
| | 4 | 0.4111 | 0.5966 | 0.2402 | 0.1037 | 0.3583 | 0.9870 | 0.0000 | 0.2154 |
| | 8 | 0.4127 | 0.5982 | 0.2544 | 0.0957 | 0.3586 | 0.9870 | 0.0000 | 0.2151 |
| | 16 | 0.4166 | 0.5986 | 0.2909 | 0.0903 | 0.3600 | 0.9870 | 0.0000 | 0.2143 |

Table C.3: nb-enron-GM1-ALL-ALL-25

| GM1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | 0.3126 | 0.4130 | 0.2686 | 0.0462 | 0.1093 | 0.8718 | 0.0000 | 0.1906 |
| | 1 | 0.3153 | 0.4779 | 0.2307 | 0.0905 | 0.2918 | 0.9157 | 0.0000 | 0.1973 |
| | 2 | 0.3191 | 0.4838 | 0.2549 | 0.0879 | 0.2950 | 0.9157 | 0.0000 | 0.1989 |
| | 4 | 0.3320 | 0.4842 | 0.2526 | 0.0846 | 0.3011 | 0.9157 | 0.0000 | 0.1974 |
| | 8 | 0.3296 | 0.4864 | 0.2641 | 0.0844 | 0.3030 | 0.9157 | 0.0000 | 0.1999 |
| | 16 | 0.3391 | 0.4875 | 0.2587 | 0.0825 | 0.3052 | 0.9157 | 0.0000 | 0.2014 |

Table C.4: nb-enron-GM1-ALL-ALL-50

| GM1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | 0.2912 | 0.3441 | 0.2603 | 0.0293 | 0.0791 | 0.8705 | 0.0000 | 0.1658 |
| | 1 | 0.2627 | 0.4048 | 0.2087 | 0.0663 | 0.2566 | 0.8085 | 0.0000 | 0.1816 |
| | 2 | 0.2800 | 0.4078 | 0.2011 | 0.0679 | 0.2625 | 0.8172 | 0.0000 | 0.1861 |
| | 4 | 0.2761 | 0.4106 | 0.2256 | 0.0626 | 0.2641 | 0.8000 | 0.0000 | 0.1848 |
| | 8 | 0.2833 | 0.4115 | 0.2237 | 0.0632 | 0.2677 | 0.8261 | 0.0000 | 0.1856 |
| | 16 | 0.2884 | 0.4136 | 0.2282 | 0.0599 | 0.2699 | 0.7917 | 0.0000 | 0.1887 |

Table C.5: nb-enron-GM1-ALL-ALL-75

| GM1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | 0.2451 | 0.2451 | 0.2450 | 0.0000 | 0.0488 | 0.8674 | 0.0000 | 0.1402 |
| | 1 | 0.1840 | 0.1841 | 0.1839 | 0.0001 | 0.1938 | 0.6728 | 0.0000 | 0.1576 |
| | 2 | 0.1898 | 0.1901 | 0.1893 | 0.0003 | 0.1971 | 0.6773 | 0.0000 | 0.1591 |
| | 4 | 0.1955 | 0.1956 | 0.1955 | 0.0001 | 0.2016 | 0.6844 | 0.0000 | 0.1604 |
| | 8 | 0.1990 | 0.1991 | 0.1989 | 0.0001 | 0.2034 | 0.6986 | 0.0000 | 0.1616 |
| | 16 | 0.2024 | 0.2028 | 0.2018 | 0.0004 | 0.2055 | 0.6926 | 0.0000 | 0.1627 |

Table C.6: nb-enron-GM1-ALL-ALL-150

| GM2 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | 0.8061 | 0.9337 | 0.5185 | 0.0740 | 0.5732 | 0.9781 | 0.0000 | 0.3272 |
| | 1 | 0.6536 | 0.8763 | 0.2766 | 0.1482 | 0.5529 | 0.9529 | 0.0000 | 0.2414 |
| | 2 | 0.7111 | 0.9132 | 0.3035 | 0.1071 | 0.5998 | 1.0000 | 0.0000 | 0.2359 |
| | 4 | 0.7320 | 0.8899 | 0.4797 | 0.0958 | 0.6136 | 0.9656 | 0.0000 | 0.2288 |
| | 8 | 0.7961 | 0.9224 | 0.5879 | 0.0753 | 0.6670 | 0.9755 | 0.0000 | 0.2165 |
| | 16 | 0.8158 | 0.9489 | 0.5926 | 0.0732 | 0.6752 | 0.9759 | 0.0000 | 0.2286 |

Table C.7: nb-enron-GM2-ALL-ALL-5

| GM2 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | 0.7209 | 0.9024 | 0.5381 | 0.0843 | 0.4862 | 0.9710 | 0.0000 | 0.3227 |
| | 1 | 0.5399 | 0.7902 | 0.2541 | 0.1209 | 0.4571 | 0.8932 | 0.0000 | 0.2309 |
| | 2 | 0.5847 | 0.7330 | 0.3271 | 0.0972 | 0.4919 | 0.9655 | 0.0000 | 0.2301 |
| | 4 | 0.6218 | 0.8022 | 0.4823 | 0.0733 | 0.5176 | 0.9241 | 0.0000 | 0.2273 |
| | 8 | 0.7130 | 0.8440 | 0.5489 | 0.0735 | 0.5794 | 0.9410 | 0.0000 | 0.2168 |
| | 16 | 0.7401 | 0.8961 | 0.5735 | 0.0780 | 0.5951 | 0.9759 | 0.0000 | 0.2308 |

Table C.8: nb-enron-GM2-ALL-ALL-10

| GM2 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | 0.6083 | 0.7145 | 0.4523 | 0.0791 | 0.3700 | 0.9737 | 0.0000 | 0.2992 |
| | 1 | 0.4166 | 0.5881 | 0.2742 | 0.0901 | 0.3675 | 0.8975 | 0.0000 | 0.2180 |
| | 2 | 0.4604 | 0.6078 | 0.3110 | 0.0756 | 0.3983 | 0.9231 | 0.0000 | 0.2158 |
| | 4 | 0.5015 | 0.5873 | 0.4101 | 0.0446 | 0.4157 | 0.9188 | 0.0000 | 0.2200 |
| | 8 | 0.6042 | 0.7139 | 0.5254 | 0.0555 | 0.4851 | 0.9257 | 0.0000 | 0.2108 |
| | 16 | 0.6469 | 0.7904 | 0.5391 | 0.0804 | 0.5119 | 0.9867 | 0.0000 | 0.2294 |

Table C.9: nb-enron-GM2-ALL-ALL-25

| GM2 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | 0.5414 | 0.5970 | 0.4515 | 0.0438 | 0.3014 | 0.9296 | 0.0000 | 0.2763 |
| | 1 | 0.3448 | 0.4884 | 0.2742 | 0.0729 | 0.3118 | 0.8958 | 0.0000 | 0.2072 |
| | 2 | 0.3831 | 0.5023 | 0.3151 | 0.0632 | 0.3371 | 0.8347 | 0.0000 | 0.2062 |
| | 4 | 0.4259 | 0.4894 | 0.3918 | 0.0335 | 0.3585 | 0.8974 | 0.0000 | 0.2119 |
| | 8 | 0.5386 | 0.5911 | 0.4971 | 0.0315 | 0.4249 | 0.8941 | 0.0000 | 0.2059 |
| | 16 | 0.5891 | 0.6972 | 0.4888 | 0.0593 | 0.4577 | 0.9589 | 0.0000 | 0.2284 |

Table C.10: nb-enron-GM2-ALL-ALL-50

| GM2 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | 0.5056 | 0.5296 | 0.4925 | 0.0127 | 0.2486 | 0.8921 | 0.0000 | 0.2603 |
| | 1 | 0.2896 | 0.4085 | 0.2295 | 0.0576 | 0.2701 | 0.8282 | 0.0000 | 0.1918 |
| | 2 | 0.3286 | 0.4265 | 0.2711 | 0.0521 | 0.3018 | 0.8235 | 0.0000 | 0.1963 |
| | 4 | 0.3762 | 0.4361 | 0.3264 | 0.0411 | 0.3246 | 0.8706 | 0.0000 | 0.2057 |
| | 8 | 0.5018 | 0.5650 | 0.4625 | 0.0392 | 0.3901 | 0.8737 | 0.0000 | 0.2020 |
| | 16 | 0.5547 | 0.6703 | 0.4654 | 0.0744 | 0.4239 | 0.9144 | 0.0000 | 0.2307 |

Table C.11: nb-enron-GM2-ALL-ALL-75

| GM2 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | 0.4536 | 0.4537 | 0.4535 | 0.0001 | 0.1706 | 0.8573 | 0.0000 | 0.2302 |
| | 1 | 0.2164 | 0.2164 | 0.2163 | 0.0000 | 0.2159 | 0.6874 | 0.0000 | 0.1776 |
| | 2 | 0.2598 | 0.2601 | 0.2593 | 0.0004 | 0.2403 | 0.7682 | 0.0000 | 0.1823 |
| | 4 | 0.3096 | 0.3097 | 0.3095 | 0.0001 | 0.2659 | 0.8385 | 0.0000 | 0.1969 |
| | 8 | 0.4547 | 0.4552 | 0.4539 | 0.0006 | 0.3333 | 0.8334 | 0.0000 | 0.2012 |
| | 16 | 0.5061 | 0.5063 | 0.5058 | 0.0002 | 0.3734 | 0.8657 | 0.0000 | 0.2351 |

Table C.12: nb-enron-GM2-ALL-ALL-150

| GM5 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | 0.7379 | 0.9618 | 0.4180 | 0.1274 | 0.5485 | 0.9870 | 0.0000 | 0.2951 |
| | 1 | 0.7817 | 0.9380 | 0.5353 | 0.0824 | 0.6598 | 0.9693 | 0.0000 | 0.2188 |
| | 2 | 0.8104 | 0.9554 | 0.6325 | 0.0755 | 0.6798 | 1.0000 | 0.0000 | 0.2241 |
| | 4 | 0.8206 | 0.9436 | 0.6265 | 0.0684 | 0.6644 | 0.9698 | 0.0000 | 0.2503 |
| | 8 | 0.8064 | 0.9372 | 0.6265 | 0.0717 | 0.6376 | 0.9718 | 0.0000 | 0.2640 |
| | 16 | 0.7980 | 0.9380 | 0.6325 | 0.0661 | 0.6032 | 0.9676 | 0.0000 | 0.2833 |

Table C.13: nb-enron-GM5-ALL-ALL-5

| GM5 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | 0.6803 | 0.9091 | 0.3708 | 0.1247 | 0.4987 | 0.9870 | 0.0000 | 0.2958 |
| | 1 | 0.6890 | 0.8903 | 0.5165 | 0.0795 | 0.5668 | 0.9505 | 0.0000 | 0.2148 |
| | 2 | 0.7274 | 0.8888 | 0.5714 | 0.0818 | 0.5972 | 0.9466 | 0.0000 | 0.2218 |
| | 4 | 0.7367 | 0.8857 | 0.5526 | 0.0816 | 0.5871 | 0.9444 | 0.0000 | 0.2447 |
| | 8 | 0.7169 | 0.8473 | 0.5485 | 0.0822 | 0.5357 | 0.9737 | 0.0000 | 0.2631 |
| | 16 | 0.6991 | 0.8389 | 0.4791 | 0.0845 | 0.5001 | 0.9867 | 0.0000 | 0.2822 |

Table C.14: nb-enron-GM5-ALL-ALL-10

| GM5 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | 0.6081 | 0.7309 | 0.4074 | 0.1104 | 0.4596 | 0.9744 | 0.0000 | 0.2962 |
| | 1 | 0.5847 | 0.7185 | 0.5079 | 0.0608 | 0.4678 | 0.9268 | 0.0000 | 0.2132 |
| | 2 | 0.6358 | 0.7188 | 0.5047 | 0.0590 | 0.5079 | 0.9620 | 0.0000 | 0.2145 |
| | 4 | 0.6445 | 0.7459 | 0.4969 | 0.0651 | 0.4997 | 0.9444 | 0.0000 | 0.2432 |
| | 8 | 0.5994 | 0.7127 | 0.5000 | 0.0450 | 0.4316 | 0.9867 | 0.0000 | 0.2539 |
| | 16 | 0.5644 | 0.6254 | 0.4921 | 0.0331 | 0.3775 | 0.9730 | 0.0000 | 0.2620 |

Table C.15: nb-enron-GM5-ALL-ALL-25

| GM5 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | 0.5742 | 0.6977 | 0.3783 | 0.0964 | 0.4284 | 0.9600 | 0.0000 | 0.3010 |
| | 1 | 0.5103 | 0.5454 | 0.4597 | 0.0311 | 0.4061 | 0.9136 | 0.0000 | 0.2113 |
| | 2 | 0.5726 | 0.6422 | 0.4907 | 0.0445 | 0.4482 | 0.9067 | 0.0000 | 0.2148 |
| | 4 | 0.5775 | 0.6355 | 0.4732 | 0.0590 | 0.4415 | 0.8986 | 0.0000 | 0.2432 |
| | 8 | 0.5142 | 0.5991 | 0.3952 | 0.0546 | 0.3477 | 0.9045 | 0.0000 | 0.2448 |
| | 16 | 0.4650 | 0.5206 | 0.3570 | 0.0466 | 0.2839 | 0.9072 | 0.0000 | 0.2414 |

Table C.16: nb-enron-GM5-ALL-ALL-50

| GM5 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | 0.5646 | 0.6346 | 0.3780 | 0.0866 | 0.4135 | 0.9444 | 0.0000 | 0.3021 |
| | 1 | 0.4722 | 0.5150 | 0.4403 | 0.0244 | 0.3675 | 0.8622 | 0.0000 | 0.2053 |
| | 2 | 0.5391 | 0.5894 | 0.4776 | 0.0370 | 0.4175 | 0.8857 | 0.0000 | 0.2106 |
| | 4 | 0.5539 | 0.5833 | 0.4791 | 0.0360 | 0.4111 | 0.8831 | 0.0000 | 0.2429 |
| | 8 | 0.4791 | 0.5279 | 0.4349 | 0.0306 | 0.3050 | 0.9046 | 0.0000 | 0.2378 |
| | 16 | 0.4316 | 0.4690 | 0.3988 | 0.0246 | 0.2359 | 0.8950 | 0.0000 | 0.2281 |

Table C.17: nb-enron-GM5-ALL-ALL-75

| GM5 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | 0.5659 | 0.5661 | 0.5657 | 0.0002 | 0.3826 | 0.9085 | 0.0000 | 0.3068 |
| | 1 | 0.4139 | 0.4141 | 0.4137 | 0.0002 | 0.3098 | 0.8095 | 0.0000 | 0.1997 |
| | 2 | 0.4941 | 0.4945 | 0.4938 | 0.0003 | 0.3648 | 0.8586 | 0.0000 | 0.2071 |
| | 4 | 0.5126 | 0.5127 | 0.5125 | 0.0001 | 0.3545 | 0.8444 | 0.0000 | 0.2491 |
| | 8 | 0.4287 | 0.4287 | 0.4285 | 0.0001 | 0.2321 | 0.8650 | 0.0000 | 0.2195 |
| | 16 | 0.2613 | 0.3703 | 0.0433 | 0.1542 | 0.1063 | 0.8920 | 0.0000 | 0.1732 |

Table C.18: nb-enron-GM5-ALL-ALL-150

| GB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | 0.7882 | 0.9709 | 0.5772 | 0.0823 | 0.5680 | 0.9852 | 0.0000 | 0.3232 |
| | 1 | 0.8167 | 0.9561 | 0.4360 | 0.1118 | 0.6987 | 0.9776 | 0.0000 | 0.2271 |
| | 2 | 0.8314 | 0.9669 | 0.4790 | 0.1051 | 0.7097 | 0.9833 | 0.0000 | 0.2332 |
| | 4 | 0.8629 | 0.9522 | 0.5556 | 0.0727 | 0.7273 | 0.9823 | 0.0000 | 0.2303 |
| | 8 | 0.8601 | 0.9782 | 0.5556 | 0.0716 | 0.7232 | 0.9889 | 0.0000 | 0.2342 |
| | 16 | 0.8589 | 0.9674 | 0.5556 | 0.0755 | 0.7174 | 1.0000 | 0.0000 | 0.2428 |

Table C.19: nb-enron-GB3-ALL-ALL-5

| GB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | 0.7057 | 0.8951 | 0.5319 | 0.0855 | 0.4602 | 0.9610 | 0.0000 | 0.3217 |
| | 1 | 0.7586 | 0.8962 | 0.5359 | 0.1064 | 0.6283 | 0.9579 | 0.0000 | 0.2354 |
| | 2 | 0.7729 | 0.9138 | 0.5161 | 0.1115 | 0.6415 | 0.9724 | 0.0000 | 0.2406 |
| | 4 | 0.8070 | 0.9251 | 0.5532 | 0.0816 | 0.6616 | 0.9688 | 0.0000 | 0.2407 |
| | 8 | 0.8074 | 0.9456 | 0.5638 | 0.0816 | 0.6570 | 1.0000 | 0.0000 | 0.2453 |
| | 16 | 0.8091 | 0.9198 | 0.5426 | 0.0873 | 0.6539 | 0.9744 | 0.0000 | 0.2535 |

Table C.20: nb-enron-GB3-ALL-ALL-10

| GB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | 0.5999 | 0.7052 | 0.4270 | 0.0712 | 0.3626 | 0.9600 | 0.0000 | 0.2942 |
| | 1 | 0.6887 | 0.8258 | 0.5542 | 0.0827 | 0.5541 | 0.9561 | 0.0000 | 0.2343 |
| | 2 | 0.7102 | 0.8493 | 0.5610 | 0.0850 | 0.5667 | 0.9620 | 0.0000 | 0.2413 |
| | 4 | 0.7520 | 0.8607 | 0.5501 | 0.0854 | 0.5903 | 0.9620 | 0.0000 | 0.2444 |
| | 8 | 0.7436 | 0.8614 | 0.5528 | 0.0834 | 0.5875 | 0.9690 | 0.0000 | 0.2452 |
| | 16 | 0.7557 | 0.8677 | 0.5556 | 0.0898 | 0.5906 | 0.9620 | 0.0000 | 0.2545 |

Table C.21: nb-enron-GB3-ALL-ALL-25

| GB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | 0.5059 | 0.5463 | 0.4256 | 0.0386 | 0.2735 | 0.9444 | 0.0000 | 0.2672 |
| | 1 | 0.6391 | 0.7451 | 0.5289 | 0.0627 | 0.5008 | 0.9394 | 0.0000 | 0.2375 |
| | 2 | 0.6740 | 0.7769 | 0.5262 | 0.0732 | 0.5194 | 0.9620 | 0.0000 | 0.2443 |
| | 4 | 0.7097 | 0.8163 | 0.5340 | 0.0834 | 0.5407 | 0.9539 | 0.0000 | 0.2486 |
| | 8 | 0.7083 | 0.7993 | 0.5255 | 0.0840 | 0.5406 | 0.9650 | 0.0000 | 0.2482 |
| | 16 | 0.7185 | 0.8277 | 0.5201 | 0.0886 | 0.5469 | 0.9744 | 0.0000 | 0.2574 |

Table C.22: nb-enron-GB3-ALL-ALL-50

| GB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | 0.4721 | 0.4962 | 0.4403 | 0.0198 | 0.2291 | 0.8811 | 0.0000 | 0.2484 |
| | 1 | 0.6188 | 0.7176 | 0.5323 | 0.0608 | 0.4748 | 0.9209 | 0.0000 | 0.2367 |
| | 2 | 0.6573 | 0.7352 | 0.5324 | 0.0684 | 0.4933 | 0.9308 | 0.0000 | 0.2453 |
| | 4 | 0.6932 | 0.7345 | 0.5359 | 0.0707 | 0.5118 | 0.9385 | 0.0000 | 0.2529 |
| | 8 | 0.6952 | 0.7491 | 0.5359 | 0.0725 | 0.5148 | 0.9615 | 0.0000 | 0.2520 |
| | 16 | 0.7075 | 0.7559 | 0.5311 | 0.0793 | 0.5192 | 0.9500 | 0.0000 | 0.2610 |

Table C.23: nb-enron-GB3-ALL-ALL-75

| GB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | 0.4282 | 0.4284 | 0.4279 | 0.0002 | 0.1611 | 0.8550 | 0.0000 | 0.2175 |
| | 1 | 0.5976 | 0.5978 | 0.5971 | 0.0004 | 0.4265 | 0.8973 | 0.0000 | 0.2451 |
| | 2 | 0.6499 | 0.6499 | 0.6499 | 0.0000 | 0.4496 | 0.9190 | 0.0000 | 0.2533 |
| | 4 | 0.6860 | 0.6862 | 0.6859 | 0.0001 | 0.4648 | 0.9348 | 0.0000 | 0.2606 |
| | 8 | 0.6889 | 0.6891 | 0.6885 | 0.0003 | 0.4701 | 0.9593 | 0.0000 | 0.2600 |
| | 16 | 0.7056 | 0.7059 | 0.7052 | 0.0003 | 0.4750 | 0.9287 | 0.0000 | 0.2688 |

Table C.24: nb-enron-GB3-ALL-ALL-150

| OSB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | 0.8527 | 0.9592 | 0.5185 | 0.0752 | 0.6957 | 1.0000 | 0.0000 | 0.2611 |
| | 1 | 0.8648 | 0.9574 | 0.5556 | 0.0741 | 0.7266 | 1.0000 | 0.0000 | 0.2416 |
| | 2 | 0.8642 | 0.9575 | 0.5556 | 0.0736 | 0.7269 | 0.9870 | 0.0000 | 0.2424 |
| | 4 | 0.8678 | 0.9587 | 0.5556 | 0.0712 | 0.7238 | 0.9870 | 0.0000 | 0.2446 |
| | 8 | 0.8638 | 0.9714 | 0.5556 | 0.0728 | 0.7260 | 1.0000 | 0.0000 | 0.2407 |
| | 16 | 0.8653 | 0.9823 | 0.5556 | 0.0732 | 0.7261 | 1.0000 | 0.0000 | 0.2417 |

Table C.25: nb-enron-OSB3-ALL-ALL-5

| OSB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | 0.7967 | 0.9336 | 0.5532 | 0.0802 | 0.6298 | 1.0000 | 0.0000 | 0.2738 |
| | 1 | 0.8161 | 0.9317 | 0.5213 | 0.0818 | 0.6584 | 0.9744 | 0.0000 | 0.2605 |
| | 2 | 0.8180 | 0.9272 | 0.5213 | 0.0844 | 0.6596 | 0.9731 | 0.0000 | 0.2606 |
| | 4 | 0.8177 | 0.9307 | 0.5213 | 0.0832 | 0.6618 | 0.9685 | 0.0000 | 0.2586 |
| | 8 | 0.8195 | 0.9332 | 0.5213 | 0.0841 | 0.6623 | 0.9744 | 0.0000 | 0.2606 |
| | 16 | 0.8186 | 0.9318 | 0.5213 | 0.0830 | 0.6590 | 0.9870 | 0.0000 | 0.2615 |

Table C.26: nb-enron-OSB3-ALL-ALL-10

| OSB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | 0.7263 | 0.8437 | 0.5711 | 0.0810 | 0.5621 | 1.0000 | 0.0000 | 0.2761 |
| | 1 | 0.7586 | 0.8489 | 0.5514 | 0.0850 | 0.5931 | 0.9870 | 0.0000 | 0.2674 |
| | 2 | 0.7635 | 0.8600 | 0.5514 | 0.0856 | 0.5964 | 0.9744 | 0.0000 | 0.2664 |
| | 4 | 0.7621 | 0.8551 | 0.5514 | 0.0858 | 0.5947 | 0.9744 | 0.0000 | 0.2644 |
| | 8 | 0.7613 | 0.8618 | 0.5556 | 0.0853 | 0.5951 | 0.9744 | 0.0000 | 0.2659 |

Table C.27: nb-enron-OSB3-ALL-ALL-25

| OSB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | 0.6818 | 0.7570 | 0.5190 | 0.0722 | 0.5105 | 1.0000 | 0.0000 | 0.2815 |
| | 1 | 0.4880 | 0.6819 | 0.3615 | 0.1021 | 0.3943 | 0.9744 | 0.0000 | 0.2598 |
| | 2 | 0.4123 | 0.5087 | 0.3526 | 0.0473 | 0.3243 | 0.9730 | 0.0000 | 0.2273 |
| | 4 | 0.7216 | 0.7899 | 0.5197 | 0.0866 | 0.5483 | 0.9744 | 0.0000 | 0.2701 |

Table C.28: nb-enron-OSB3-ALL-ALL-50

| OSB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | 0.6713 | 0.7004 | 0.5504 | 0.0542 | 0.4870 | 1.0000 | 0.0000 | 0.2883 |
| | 1 | 0.7074 | 0.7478 | 0.5376 | 0.0762 | 0.5221 | 0.9287 | 0.0000 | 0.2748 |
| | 2 | 0.7112 | 0.7531 | 0.5390 | 0.0771 | 0.5218 | 0.9341 | 0.0000 | 0.2757 |
| | 4 | 0.7089 | 0.7621 | 0.5393 | 0.0773 | 0.5199 | 0.9352 | 0.0000 | 0.2757 |
| | 8 | 0.7079 | 0.7860 | 0.5389 | 0.0829 | 0.5230 | 0.9290 | 0.0000 | 0.2750 |

Table C.29: nb-enron-OSB3-ALL-ALL-75

| OSB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | 0.6572 | 0.6574 | 0.6571 | 0.0002 | 0.4441 | 0.9867 | 0.0000 | 0.3023 |
| | 1 | 0.7041 | 0.7042 | 0.7040 | 0.0001 | 0.4789 | 0.9127 | 0.0000 | 0.2855 |
| | 2 | 0.7091 | 0.7092 | 0.7089 | 0.0001 | 0.4787 | 0.9146 | 0.0000 | 0.2875 |
| | 4 | 0.7068 | 0.7071 | 0.7066 | 0.0003 | 0.4769 | 0.9189 | 0.0000 | 0.2874 |
| | 8 | 0.7101 | 0.7101 | 0.7100 | 0.0001 | 0.4778 | 0.9151 | 0.0000 | 0.2878 |

Table C.30: nb-enron-OSB3-ALL-ALL-150

# APPENDIX D:
# Naive Bayes Accuracy and F-Score Results for the Twitter Short Message Corpus

| GM1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | 0.6264 | 0.7714 | 0.4580 | 0.0794 | 0.5886 | 0.9618 | 0.0000 | 0.1854 |
| | 1 | 0.4974 | 0.7500 | 0.1965 | 0.1176 | 0.4734 | 0.9421 | 0.0000 | 0.1630 |
| | 2 | 0.5130 | 0.8101 | 0.2785 | 0.1138 | 0.4936 | 0.9463 | 0.0615 | 0.1538 |
| | 4 | 0.5132 | 0.7749 | 0.2637 | 0.1029 | 0.4914 | 0.9518 | 0.1190 | 0.1537 |
| | 8 | 0.5238 | 0.7913 | 0.2861 | 0.1088 | 0.5019 | 0.9562 | 0.1311 | 0.1531 |
| | 16 | 0.5102 | 0.7778 | 0.2885 | 0.1105 | 0.4904 | 0.9501 | 0.1649 | 0.1509 |

Table D.1: nb-twitter-GM1-ALL-ALL-5

| GM1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | 0.4869 | 0.6411 | 0.2738 | 0.0831 | 0.4345 | 0.9282 | 0.0000 | 0.2191 |
| | 1 | 0.3490 | 0.5598 | 0.1297 | 0.1163 | 0.3227 | 0.9152 | 0.0000 | 0.1697 |
| | 2 | 0.3671 | 0.6005 | 0.1623 | 0.1142 | 0.3407 | 0.9209 | 0.0000 | 0.1662 |
| | 4 | 0.3693 | 0.5990 | 0.1997 | 0.1045 | 0.3437 | 0.9151 | 0.0274 | 0.1600 |
| | 8 | 0.3755 | 0.5902 | 0.1692 | 0.1023 | 0.3498 | 0.9002 | 0.0000 | 0.1644 |
| | 16 | 0.3642 | 0.5588 | 0.1639 | 0.1141 | 0.3398 | 0.8667 | 0.0000 | 0.1616 |

Table D.2: nb-twitter-GM1-ALL-ALL-10

| GM1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | 0.3357 | 0.4389 | 0.2200 | 0.0642 | 0.2848 | 0.8560 | 0.0000 | 0.2063 |
| | 1 | 0.2260 | 0.3653 | 0.0731 | 0.0991 | 0.2007 | 0.8185 | 0.0000 | 0.1561 |
| | 2 | 0.2299 | 0.3574 | 0.0817 | 0.0963 | 0.2042 | 0.8385 | 0.0000 | 0.1556 |
| | 4 | 0.2332 | 0.3539 | 0.0956 | 0.0918 | 0.2088 | 0.8615 | 0.0000 | 0.1559 |
| | 8 | 0.2398 | 0.3882 | 0.0848 | 0.0934 | 0.2132 | 0.8615 | 0.0000 | 0.1571 |
| | 16 | 0.2304 | 0.3517 | 0.0733 | 0.0965 | 0.2023 | 0.8803 | 0.0000 | 0.1535 |

Table D.3: nb-twitter-GM1-ALL-ALL-25

| GM1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | 0.2326 | 0.3191 | 0.1395 | 0.0509 | 0.1920 | 0.8426 | 0.0000 | 0.1850 |
| | 1 | 0.1595 | 0.2449 | 0.0387 | 0.0799 | 0.1323 | 0.8116 | 0.0000 | 0.1393 |
| | 2 | 0.1631 | 0.2501 | 0.0460 | 0.0790 | 0.1366 | 0.8615 | 0.0000 | 0.1368 |
| | 4 | 0.1653 | 0.2507 | 0.0556 | 0.0779 | 0.1404 | 0.8000 | 0.0000 | 0.1355 |
| | 8 | 0.1671 | 0.2522 | 0.0504 | 0.0757 | 0.1418 | 0.8750 | 0.0000 | 0.1409 |
| | 16 | 0.1638 | 0.2436 | 0.0486 | 0.0807 | 0.1387 | 0.8485 | 0.0000 | 0.1426 |

Table D.4: nb-twitter-GM1-ALL-ALL-50

| GM1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | 0.1820 | 0.2356 | 0.1399 | 0.0310 | 0.1439 | 0.7876 | 0.0000 | 0.1695 |
| | 1 | 0.1328 | 0.2014 | 0.0254 | 0.0707 | 0.1070 | 0.7324 | 0.0000 | 0.1289 |
| | 2 | 0.1364 | 0.2167 | 0.0303 | 0.0723 | 0.1115 | 0.8000 | 0.0000 | 0.1310 |
| | 4 | 0.1370 | 0.2184 | 0.0400 | 0.0691 | 0.1132 | 0.7941 | 0.0000 | 0.1283 |
| | 8 | 0.1390 | 0.2275 | 0.0404 | 0.0698 | 0.1156 | 0.7617 | 0.0000 | 0.1309 |
| | 16 | 0.1360 | 0.2114 | 0.0315 | 0.0731 | 0.1109 | 0.7680 | 0.0000 | 0.1268 |

Table D.5: nb-twitter-GM1-ALL-ALL-75

| GM1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | 0.1252 | 0.1254 | 0.1247 | 0.0003 | 0.0855 | 0.7565 | 0.0000 | 0.1448 |
| | 1 | 0.0982 | 0.1353 | 0.0239 | 0.0525 | 0.0745 | 0.7500 | 0.0000 | 0.1112 |
| | 2 | 0.0979 | 0.1344 | 0.0248 | 0.0517 | 0.0738 | 0.7027 | 0.0000 | 0.1102 |
| | 4 | 0.0975 | 0.1351 | 0.0231 | 0.0526 | 0.0760 | 0.7324 | 0.0000 | 0.1091 |
| | 8 | 0.0986 | 0.1350 | 0.0274 | 0.0504 | 0.0754 | 0.7123 | 0.0000 | 0.1101 |
| | 16 | 0.0983 | 0.1386 | 0.0200 | 0.0554 | 0.0758 | 0.6923 | 0.0000 | 0.1095 |

Table D.6: nb-twitter-GM1-ALL-ALL-150

| GM2 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | 0.5711 | 0.7890 | 0.3875 | 0.0853 | 0.5299 | 0.9486 | 0.0000 | 0.1822 |
| | 1 | 0.5253 | 0.7643 | 0.2871 | 0.1167 | 0.5135 | 0.9250 | 0.1667 | 0.1530 |
| | 2 | 0.5301 | 0.7467 | 0.3114 | 0.1090 | 0.5175 | 0.9328 | 0.1284 | 0.1495 |
| | 4 | 0.5401 | 0.7926 | 0.3519 | 0.0979 | 0.5259 | 0.9237 | 0.1522 | 0.1468 |
| | 8 | 0.5459 | 0.7992 | 0.2906 | 0.1085 | 0.5313 | 0.9457 | 0.1944 | 0.1512 |
| | 16 | 0.5409 | 0.8437 | 0.2975 | 0.1141 | 0.5249 | 0.9560 | 0.1481 | 0.1537 |

Table D.7: nb-twitter-GM2-ALL-ALL-5

| GM2 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | 0.4439 | 0.6304 | 0.2809 | 0.0749 | 0.4032 | 0.8824 | 0.0000 | 0.1885 |
| | 1 | 0.3847 | 0.5960 | 0.1816 | 0.1107 | 0.3716 | 0.8974 | 0.0377 | 0.1667 |
| | 2 | 0.3914 | 0.5446 | 0.1842 | 0.1064 | 0.3781 | 0.8988 | 0.0519 | 0.1705 |
| | 4 | 0.4018 | 0.6610 | 0.2146 | 0.1041 | 0.3879 | 0.8947 | 0.0588 | 0.1657 |
| | 8 | 0.4046 | 0.6051 | 0.1764 | 0.1088 | 0.3866 | 0.9231 | 0.0250 | 0.1688 |
| | 16 | 0.4031 | 0.6179 | 0.1855 | 0.1147 | 0.3832 | 0.9170 | 0.0357 | 0.1759 |

Table D.8: nb-twitter-GM2-ALL-ALL-10

| GM2 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | 0.3215 | 0.3855 | 0.2617 | 0.0382 | 0.2821 | 0.8462 | 0.0000 | 0.1846 |
| | 1 | 0.2508 | 0.4249 | 0.0872 | 0.1062 | 0.2406 | 0.8755 | 0.0000 | 0.1721 |
| | 2 | 0.2510 | 0.3846 | 0.0963 | 0.1036 | 0.2406 | 0.8745 | 0.0000 | 0.1694 |
| | 4 | 0.2686 | 0.4173 | 0.1121 | 0.0960 | 0.2559 | 0.8364 | 0.0000 | 0.1681 |
| | 8 | 0.2768 | 0.4602 | 0.1198 | 0.1051 | 0.2597 | 0.8681 | 0.0000 | 0.1714 |
| | 16 | 0.2703 | 0.4437 | 0.1105 | 0.1007 | 0.2545 | 0.8619 | 0.0000 | 0.1747 |

Table D.9: nb-twitter-GM2-ALL-ALL-25

| GM2 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | 0.2509 | 0.2992 | 0.2038 | 0.0282 | 0.2175 | 0.8438 | 0.0000 | 0.1705 |
| | 1 | 0.1854 | 0.3044 | 0.0507 | 0.0938 | 0.1758 | 0.8288 | 0.0000 | 0.1584 |
| | 2 | 0.1843 | 0.2660 | 0.0413 | 0.0909 | 0.1743 | 0.8300 | 0.0000 | 0.1629 |
| | 4 | 0.1980 | 0.3055 | 0.0578 | 0.0901 | 0.1844 | 0.8037 | 0.0000 | 0.1614 |
| | 8 | 0.2058 | 0.2880 | 0.0771 | 0.0883 | 0.1919 | 0.8362 | 0.0000 | 0.1619 |
| | 16 | 0.2023 | 0.3031 | 0.0786 | 0.0892 | 0.1891 | 0.8438 | 0.0000 | 0.1662 |

Table D.10: nb-twitter-GM2-ALL-ALL-50

| GM2 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | 0.2162 | 0.2621 | 0.1630 | 0.0338 | 0.1846 | 0.7879 | 0.0000 | 0.1622 |
| | 1 | 0.1518 | 0.2358 | 0.0368 | 0.0823 | 0.1431 | 0.8036 | 0.0000 | 0.1514 |
| | 2 | 0.1535 | 0.2255 | 0.0328 | 0.0831 | 0.1448 | 0.8293 | 0.0000 | 0.1501 |
| | 4 | 0.1666 | 0.2477 | 0.0477 | 0.0835 | 0.1550 | 0.8073 | 0.0000 | 0.1516 |
| | 8 | 0.1770 | 0.2535 | 0.0578 | 0.0825 | 0.1648 | 0.8106 | 0.0000 | 0.1596 |
| | 16 | 0.1700 | 0.2579 | 0.0573 | 0.0808 | 0.1573 | 0.8571 | 0.0000 | 0.1580 |

Table D.11: nb-twitter-GM2-ALL-ALL-75

| GM2 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | 0.1709 | 0.1719 | 0.1690 | 0.0013 | 0.1379 | 0.7619 | 0.0000 | 0.1484 |
| | 1 | 0.1155 | 0.1655 | 0.0190 | 0.0683 | 0.1074 | 0.7593 | 0.0000 | 0.1354 |
| | 2 | 0.1173 | 0.1659 | 0.0206 | 0.0684 | 0.1104 | 0.7967 | 0.0000 | 0.1397 |
| | 4 | 0.1282 | 0.1789 | 0.0276 | 0.0712 | 0.1169 | 0.7477 | 0.0000 | 0.1385 |
| | 8 | 0.1331 | 0.1819 | 0.0363 | 0.0684 | 0.1236 | 0.7414 | 0.0000 | 0.1437 |
| | 16 | 0.1290 | 0.1760 | 0.0357 | 0.0660 | 0.1169 | 0.7519 | 0.0000 | 0.1425 |

Table D.12: nb-twitter-GM2-ALL-ALL-150

| GM5 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | 0.3453 | 0.5293 | 0.2431 | 0.0531 | 0.2306 | 0.7931 | 0.0000 | 0.1930 |
| | 1 | 0.5530 | 0.7495 | 0.4066 | 0.0794 | 0.5324 | 0.9106 | 0.1096 | 0.1415 |
| | 2 | 0.5523 | 0.7220 | 0.4000 | 0.0789 | 0.5279 | 0.9231 | 0.1212 | 0.1530 |
| | 4 | 0.5516 | 0.7077 | 0.3814 | 0.0714 | 0.5223 | 0.9052 | 0.0426 | 0.1564 |
| | 8 | 0.5550 | 0.7094 | 0.3968 | 0.0687 | 0.5262 | 0.9254 | 0.0779 | 0.1559 |
| | 16 | 0.5579 | 0.7680 | 0.4106 | 0.0777 | 0.5299 | 0.9245 | 0.0571 | 0.1643 |

Table D.13: nb-twitter-GM5-ALL-ALL-5

| GM5 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | 0.2408 | 0.3357 | 0.1748 | 0.0354 | 0.1711 | 0.8136 | 0.0000 | 0.1595 |
| | 1 | 0.4222 | 0.5466 | 0.3065 | 0.0630 | 0.3994 | 0.8571 | 0.0556 | 0.1545 |
| | 2 | 0.4187 | 0.5806 | 0.3035 | 0.0642 | 0.3954 | 0.8889 | 0.0267 | 0.1652 |
| | 4 | 0.4191 | 0.5954 | 0.3069 | 0.0590 | 0.3913 | 0.8618 | 0.0286 | 0.1711 |
| | 8 | 0.4209 | 0.5239 | 0.3269 | 0.0506 | 0.3953 | 0.8750 | 0.0303 | 0.1673 |
| | 16 | 0.4319 | 0.5606 | 0.3228 | 0.0590 | 0.4062 | 0.8932 | 0.0267 | 0.1710 |

Table D.14: nb-twitter-GM5-ALL-ALL-10

| GM5 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | 0.1543 | 0.2141 | 0.1224 | 0.0232 | 0.1288 | 0.7931 | 0.0000 | 0.1443 |
| | 1 | 0.2990 | 0.3613 | 0.2514 | 0.0302 | 0.2770 | 0.8190 | 0.0000 | 0.1613 |
| | 2 | 0.2923 | 0.3493 | 0.2405 | 0.0324 | 0.2728 | 0.8710 | 0.0000 | 0.1661 |
| | 4 | 0.2935 | 0.3300 | 0.2133 | 0.0306 | 0.2689 | 0.8750 | 0.0000 | 0.1720 |
| | 8 | 0.2929 | 0.3552 | 0.2278 | 0.0293 | 0.2722 | 0.8438 | 0.0000 | 0.1707 |
| | 16 | 0.2973 | 0.3765 | 0.2032 | 0.0395 | 0.2768 | 0.8615 | 0.0000 | 0.1740 |

Table D.15: nb-twitter-GM5-ALL-ALL-25

| GM5 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | 0.1226 | 0.1558 | 0.0999 | 0.0157 | 0.1079 | 0.7719 | 0.0000 | 0.1390 |
| | 1 | 0.2277 | 0.2534 | 0.1808 | 0.0276 | 0.2096 | 0.7961 | 0.0000 | 0.1528 |
| | 2 | 0.2252 | 0.2637 | 0.1996 | 0.0205 | 0.2084 | 0.8000 | 0.0000 | 0.1630 |
| | 4 | 0.2216 | 0.2441 | 0.1770 | 0.0182 | 0.2031 | 0.8254 | 0.0000 | 0.1658 |
| | 8 | 0.2217 | 0.2573 | 0.1871 | 0.0212 | 0.2033 | 0.8224 | 0.0000 | 0.1637 |
| | 16 | 0.2245 | 0.2608 | 0.1540 | 0.0308 | 0.2059 | 0.8710 | 0.0000 | 0.1644 |

Table D.16: nb-twitter-GM5-ALL-ALL-50

| GM5 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | 0.0977 | 0.1069 | 0.0836 | 0.0086 | 0.0908 | 0.7719 | 0.0000 | 0.1280 |
| | 1 | 0.1960 | 0.2111 | 0.1780 | 0.0121 | 0.1782 | 0.6739 | 0.0000 | 0.1457 |
| | 2 | 0.1952 | 0.2113 | 0.1803 | 0.0122 | 0.1767 | 0.8710 | 0.0000 | 0.1520 |
| | 4 | 0.1878 | 0.1984 | 0.1777 | 0.0069 | 0.1725 | 0.7937 | 0.0000 | 0.1565 |
| | 8 | 0.1902 | 0.1945 | 0.1861 | 0.0029 | 0.1729 | 0.8065 | 0.0000 | 0.1565 |
| | 16 | 0.1908 | 0.2060 | 0.1689 | 0.0135 | 0.1741 | 0.7937 | 0.0000 | 0.1585 |

Table D.17: nb-twitter-GM5-ALL-ALL-75

| GM5 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | 0.0787 | 0.0792 | 0.0784 | 0.0004 | 0.0754 | 0.6429 | 0.0000 | 0.1186 |
| | 1 | 0.1535 | 0.1542 | 0.1531 | 0.0005 | 0.1370 | 0.6392 | 0.0000 | 0.1359 |
| | 2 | 0.1524 | 0.1531 | 0.1520 | 0.0005 | 0.1360 | 0.7879 | 0.0000 | 0.1427 |
| | 4 | 0.1404 | 0.1414 | 0.1399 | 0.0007 | 0.1276 | 0.7500 | 0.0000 | 0.1459 |
| | 8 | 0.1401 | 0.1404 | 0.1399 | 0.0002 | 0.1267 | 0.7463 | 0.0000 | 0.1443 |
| | 16 | 0.1424 | 0.1430 | 0.1420 | 0.0005 | 0.1294 | 0.7813 | 0.0000 | 0.1465 |

Table D.18: nb-twitter-GM5-ALL-ALL-150

| GB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | 0.6179 | 0.8021 | 0.4745 | 0.0729 | 0.5813 | 0.9284 | 0.0392 | 0.1654 |
| | 1 | 0.6109 | 0.8125 | 0.4585 | 0.0734 | 0.5959 | 0.9480 | 0.2000 | 0.1296 |
| | 2 | 0.6160 | 0.7823 | 0.4606 | 0.0699 | 0.5986 | 0.9347 | 0.2340 | 0.1291 |
| | 4 | 0.6199 | 0.8546 | 0.4694 | 0.0819 | 0.6043 | 0.9613 | 0.2444 | 0.1330 |
| | 8 | 0.6187 | 0.8474 | 0.4669 | 0.0770 | 0.6040 | 0.9409 | 0.2222 | 0.1266 |
| | 16 | 0.6265 | 0.8216 | 0.4648 | 0.0773 | 0.6098 | 0.9512 | 0.2154 | 0.1322 |

Table D.19: nb-twitter-GB3-ALL-ALL-5

| GB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | 0.4948 | 0.6108 | 0.3477 | 0.0607 | 0.4543 | 0.9091 | 0.0000 | 0.1799 |
| | 1 | 0.4975 | 0.6356 | 0.3297 | 0.0644 | 0.4760 | 0.9002 | 0.1190 | 0.1515 |
| | 2 | 0.5015 | 0.7366 | 0.3748 | 0.0712 | 0.4809 | 0.9289 | 0.1075 | 0.1444 |
| | 4 | 0.5011 | 0.6671 | 0.3665 | 0.0723 | 0.4801 | 0.9102 | 0.0741 | 0.1521 |
| | 8 | 0.4999 | 0.6359 | 0.3732 | 0.0716 | 0.4801 | 0.9197 | 0.0952 | 0.1489 |
| | 16 | 0.5041 | 0.6385 | 0.3876 | 0.0636 | 0.4838 | 0.8986 | 0.0909 | 0.1465 |

Table D.20: nb-twitter-GB3-ALL-ALL-10

| GB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | 0.3584 | 0.4401 | 0.2988 | 0.0397 | 0.3164 | 0.8696 | 0.0000 | 0.1861 |
| | 1 | 0.3724 | 0.4429 | 0.3147 | 0.0367 | 0.3490 | 0.8824 | 0.0202 | 0.1557 |
| | 2 | 0.3760 | 0.4394 | 0.3271 | 0.0352 | 0.3530 | 0.8155 | 0.0227 | 0.1507 |
| | 4 | 0.3773 | 0.4834 | 0.3207 | 0.0466 | 0.3539 | 0.8504 | 0.0227 | 0.1554 |
| | 8 | 0.3733 | 0.4633 | 0.3186 | 0.0447 | 0.3528 | 0.8355 | 0.0244 | 0.1524 |
| | 16 | 0.3838 | 0.4431 | 0.2890 | 0.0417 | 0.3600 | 0.8649 | 0.0270 | 0.1619 |

Table D.21: nb-twitter-GB3-ALL-ALL-25

| GB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | 0.2839 | 0.3292 | 0.2411 | 0.0265 | 0.2457 | 0.8955 | 0.0000 | 0.1769 |
| | 1 | 0.2984 | 0.3356 | 0.2686 | 0.0234 | 0.2746 | 0.8219 | 0.0000 | 0.1569 |
| | 2 | 0.3019 | 0.3366 | 0.2770 | 0.0222 | 0.2778 | 0.7529 | 0.0000 | 0.1500 |
| | 4 | 0.3044 | 0.3376 | 0.2728 | 0.0237 | 0.2803 | 0.7895 | 0.0000 | 0.1540 |
| | 8 | 0.3038 | 0.3409 | 0.2618 | 0.0312 | 0.2805 | 0.8067 | 0.0000 | 0.1536 |
| | 16 | 0.3100 | 0.3411 | 0.2813 | 0.0236 | 0.2858 | 0.8696 | 0.0000 | 0.1595 |

Table D.22: nb-twitter-GB3-ALL-ALL-50

| GB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | 0.2484 | 0.2862 | 0.2169 | 0.0247 | 0.2120 | 0.8615 | 0.0000 | 0.1667 |
| | 1 | 0.2648 | 0.3021 | 0.2254 | 0.0309 | 0.2414 | 0.8000 | 0.0000 | 0.1551 |
| | 2 | 0.2668 | 0.3023 | 0.2297 | 0.0295 | 0.2423 | 0.7333 | 0.0000 | 0.1481 |
| | 4 | 0.2667 | 0.3128 | 0.2263 | 0.0300 | 0.2433 | 0.7368 | 0.0000 | 0.1541 |
| | 8 | 0.2657 | 0.2990 | 0.2368 | 0.0235 | 0.2434 | 0.7478 | 0.0171 | 0.1481 |
| | 16 | 0.2748 | 0.3096 | 0.2389 | 0.0259 | 0.2512 | 0.8116 | 0.0000 | 0.1578 |

Table D.23: nb-twitter-GB3-ALL-ALL-75

| GB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | 0.1988 | 0.1995 | 0.1973 | 0.0011 | 0.1613 | 0.7879 | 0.0000 | 0.1550 |
| | 1 | 0.2152 | 0.2154 | 0.2148 | 0.0003 | 0.1910 | 0.7105 | 0.0000 | 0.1454 |
| | 2 | 0.2190 | 0.2191 | 0.2186 | 0.0002 | 0.1945 | 0.6458 | 0.0000 | 0.1397 |
| | 4 | 0.2197 | 0.2200 | 0.2190 | 0.0005 | 0.1954 | 0.6914 | 0.0000 | 0.1450 |
| | 8 | 0.2180 | 0.2187 | 0.2176 | 0.0006 | 0.1942 | 0.7000 | 0.0000 | 0.1418 |
| | 16 | 0.2221 | 0.2233 | 0.2216 | 0.0008 | 0.1976 | 0.7297 | 0.0000 | 0.1476 |

Table D.24: nb-twitter-GB3-ALL-ALL-150

| OSB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | 0.6475 | 0.8164 | 0.4983 | 0.0673 | 0.6308 | 0.9254 | 0.1481 | 0.1296 |
| | 1 | 0.5628 | 0.8293 | 0.3913 | 0.0878 | 0.5296 | 0.9419 | 0.0597 | 0.1640 |
| | 2 | 0.5687 | 0.7734 | 0.3836 | 0.0833 | 0.5328 | 0.9243 | 0.0597 | 0.1675 |
| | 4 | 0.5627 | 0.7768 | 0.3836 | 0.0856 | 0.5292 | 0.9419 | 0.0597 | 0.1656 |
| | 8 | 0.5626 | 0.7790 | 0.3429 | 0.0915 | 0.5272 | 0.9419 | 0.0597 | 0.1677 |
| | 16 | 0.5587 | 0.7479 | 0.3857 | 0.0802 | 0.5251 | 0.9458 | 0.0351 | 0.1636 |

Table D.25: nb-twitter-OSB3-ALL-ALL-5

| OSB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | 0.5271 | 0.6410 | 0.4333 | 0.0545 | 0.5077 | 0.9069 | 0.0988 | 0.1454 |
| | 1 | 0.4387 | 0.6071 | 0.3160 | 0.0682 | 0.3977 | 0.9122 | 0.0000 | 0.1724 |
| | 2 | 0.4420 | 0.5840 | 0.2883 | 0.0688 | 0.4002 | 0.9098 | 0.0000 | 0.1752 |
| | 4 | 0.4403 | 0.6238 | 0.2876 | 0.0767 | 0.3971 | 0.9228 | 0.0000 | 0.1777 |
| | 8 | 0.4410 | 0.5966 | 0.2853 | 0.0690 | 0.3983 | 0.9228 | 0.0000 | 0.1771 |
| | 16 | 0.4425 | 0.5954 | 0.3245 | 0.0614 | 0.3991 | 0.9265 | 0.0000 | 0.1757 |

Table D.26: nb-twitter-OSB3-ALL-ALL-10

| OSB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | 0.3979 | 0.4848 | 0.3207 | 0.0464 | 0.3752 | 0.8824 | 0.0000 | 0.1648 |
| | 1 | 0.3206 | 0.4237 | 0.2419 | 0.0407 | 0.2744 | 0.8341 | 0.0000 | 0.1689 |
| | 2 | 0.3223 | 0.3951 | 0.2704 | 0.0365 | 0.2776 | 0.8386 | 0.0000 | 0.1695 |
| | 4 | 0.3228 | 0.3814 | 0.2773 | 0.0359 | 0.2759 | 0.8571 | 0.0000 | 0.1679 |
| | 8 | 0.3249 | 0.3937 | 0.2570 | 0.0369 | 0.2814 | 0.8389 | 0.0000 | 0.1682 |
| | 16 | 0.3229 | 0.4016 | 0.2749 | 0.0351 | 0.2785 | 0.8629 | 0.0000 | 0.1698 |

Table D.27: nb-twitter-OSB3-ALL-ALL-25

| OSB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | 0.3233 | 0.3739 | 0.2838 | 0.0259 | 0.3004 | 0.7895 | 0.0000 | 0.1641 |
| | 1 | 0.2535 | 0.2797 | 0.2254 | 0.0177 | 0.2115 | 0.7556 | 0.0000 | 0.1573 |
| | 2 | 0.2538 | 0.2810 | 0.2332 | 0.0177 | 0.2114 | 0.7712 | 0.0000 | 0.1569 |
| | 4 | 0.2548 | 0.2819 | 0.2296 | 0.0181 | 0.2126 | 0.7511 | 0.0000 | 0.1594 |
| | 8 | 0.2561 | 0.2831 | 0.2305 | 0.0200 | 0.2144 | 0.7585 | 0.0000 | 0.1587 |
| | 16 | 0.2564 | 0.2913 | 0.2300 | 0.0216 | 0.2136 | 0.7609 | 0.0000 | 0.1594 |

Table D.28: nb-twitter-OSB3-ALL-ALL-50

| OSB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | 0.2861 | 0.3048 | 0.2582 | 0.0180 | 0.2634 | 0.7945 | 0.0000 | 0.1605 |
| | 1 | 0.2212 | 0.2444 | 0.1972 | 0.0198 | 0.1818 | 0.7364 | 0.0000 | 0.1513 |
| | 2 | 0.2242 | 0.2654 | 0.1806 | 0.0267 | 0.1840 | 0.8051 | 0.0000 | 0.1513 |
| | 4 | 0.2237 | 0.2615 | 0.1851 | 0.0269 | 0.1851 | 0.7229 | 0.0000 | 0.1512 |
| | 8 | 0.2219 | 0.2611 | 0.1870 | 0.0237 | 0.1831 | 0.7077 | 0.0000 | 0.1500 |
| | 16 | 0.2228 | 0.2482 | 0.1973 | 0.0209 | 0.1838 | 0.7360 | 0.0000 | 0.1515 |

Table D.29: nb-twitter-OSB3-ALL-ALL-75

| OSB3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Accuracy | | | | F-Score | | | |
| | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | 0.2332 | 0.2337 | 0.2320 | 0.0008 | 0.2104 | 0.7778 | 0.0000 | 0.1561 |
| | 1 | 0.1775 | 0.1776 | 0.1773 | 0.0001 | 0.1402 | 0.6337 | 0.0000 | 0.1314 |
| | 2 | 0.1791 | 0.1792 | 0.1790 | 0.0001 | 0.1422 | 0.6250 | 0.0000 | 0.1322 |
| | 4 | 0.1786 | 0.1788 | 0.1784 | 0.0002 | 0.1412 | 0.6244 | 0.0000 | 0.1329 |
| | 8 | 0.1795 | 0.1810 | 0.1787 | 0.0011 | 0.1422 | 0.6570 | 0.0000 | 0.1328 |
| | 16 | 0.1789 | 0.1797 | 0.1785 | 0.0006 | 0.1411 | 0.6540 | 0.0000 | 0.1329 |

Table D.30: nb-twitter-OSB3-ALL-ALL-150

# APPENDIX E: Grouped Results LibLinear Results for the ENRON Email Corpus

| GM1 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | Group Type | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | RAN | 0.8581 | 0.9505 | 0.6630 | 0.0607 | 0.6718 | 0.9795 | 0.0000 | 0.2650 |
| | | SAL | 0.8767 | 0.9531 | 0.7526 | 0.0517 | 0.6663 | 0.9842 | 0.0000 | 0.2625 |
| | | STL | 0.7460 | 0.9246 | 0.4815 | 0.1113 | 0.7211 | 0.9653 | 0.0000 | 0.1841 |
| | 1 | RAN | 0.8475 | 0.9362 | 0.6603 | 0.0762 | 0.6682 | 0.9737 | 0.0000 | 0.2516 |
| | | SAL | 0.8797 | 0.9578 | 0.7977 | 0.0363 | 0.6682 | 0.9826 | 0.0000 | 0.2670 |
| | | STL | 0.7426 | 0.9332 | 0.4444 | 0.1129 | 0.7215 | 0.9712 | 0.0000 | 0.1845 |
| | 2 | RAN | 0.8400 | 0.9275 | 0.5742 | 0.0747 | 0.6716 | 0.9672 | 0.0000 | 0.2570 |
| | | SAL | 0.8810 | 0.9570 | 0.8049 | 0.0337 | 0.6701 | 0.9819 | 0.0000 | 0.2621 |
| | | STL | 0.7437 | 0.9291 | 0.4444 | 0.1080 | 0.7226 | 0.9692 | 0.0000 | 0.1817 |
| | 4 | RAN | 0.8639 | 0.9590 | 0.7444 | 0.0553 | 0.6833 | 0.9799 | 0.0000 | 0.2500 |
| | | SAL | 0.8800 | 0.9582 | 0.7877 | 0.0365 | 0.6766 | 0.9821 | 0.0000 | 0.2535 |
| | | STL | 0.7455 | 0.9314 | 0.4444 | 0.1089 | 0.7250 | 0.9676 | 0.0000 | 0.1807 |
| | 8 | RAN | 0.8656 | 0.9485 | 0.7321 | 0.0656 | 0.6705 | 0.9760 | 0.0000 | 0.2671 |
| | | SAL | 0.8810 | 0.9570 | 0.7848 | 0.0347 | 0.6720 | 0.9819 | 0.0000 | 0.2592 |
| | | STL | 0.7427 | 0.9237 | 0.4444 | 0.1087 | 0.7208 | 0.9651 | 0.0000 | 0.1829 |
| | 16 | RAN | 0.8550 | 0.9555 | 0.6913 | 0.0616 | 0.6795 | 0.9744 | 0.0000 | 0.2492 |
| | | SAL | 0.8789 | 0.9732 | 0.7844 | 0.0369 | 0.6731 | 0.9878 | 0.0000 | 0.2536 |
| | | STL | 0.7379 | 0.9269 | 0.4444 | 0.1127 | 0.7178 | 0.9668 | 0.0000 | 0.1827 |

Table E.1: grouped-liblinear-enron-GM1-ALL-ALL-5

| GM1 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | Group Type | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | RAN | 0.8024 | 0.9312 | 0.6768 | 0.0655 | 0.5981 | 0.9778 | 0.0000 | 0.2636 |
| | | SAL | 0.8227 | 0.9043 | 0.7375 | 0.0531 | 0.6038 | 0.9660 | 0.0000 | 0.2671 |
| | | STL | 0.6582 | 0.8234 | 0.3776 | 0.1228 | 0.6346 | 0.9618 | 0.0000 | 0.2012 |
| | 1 | RAN | 0.8123 | 0.8862 | 0.7168 | 0.0572 | 0.6040 | 0.9699 | 0.0000 | 0.2686 |
| | | SAL | 0.8209 | 0.8890 | 0.7641 | 0.0360 | 0.5922 | 0.9620 | 0.0000 | 0.2714 |
| | | STL | 0.6499 | 0.8117 | 0.3878 | 0.1172 | 0.6282 | 0.9397 | 0.0000 | 0.1990 |
| | 2 | RAN | 0.7988 | 0.9068 | 0.6456 | 0.0668 | 0.5900 | 0.9660 | 0.0000 | 0.2655 |
| | | SAL | 0.8246 | 0.8905 | 0.7788 | 0.0331 | 0.6025 | 0.9633 | 0.0000 | 0.2633 |
| | | STL | 0.6547 | 0.8168 | 0.4388 | 0.1124 | 0.6353 | 0.9488 | 0.0000 | 0.1932 |
| | 4 | RAN | 0.8083 | 0.9086 | 0.6986 | 0.0481 | 0.6013 | 0.9692 | 0.0000 | 0.2674 |
| | | SAL | 0.8221 | 0.8922 | 0.7372 | 0.0367 | 0.5968 | 0.9652 | 0.0000 | 0.2651 |
| | | STL | 0.6502 | 0.8177 | 0.4388 | 0.1127 | 0.6299 | 0.9428 | 0.0000 | 0.1943 |
| | 8 | RAN | 0.8027 | 0.8736 | 0.6660 | 0.0616 | 0.6055 | 0.9678 | 0.0000 | 0.2555 |
| | | SAL | 0.8182 | 0.9025 | 0.7224 | 0.0426 | 0.5966 | 0.9684 | 0.0000 | 0.2683 |
| | | STL | 0.6525 | 0.8144 | 0.4388 | 0.1125 | 0.6319 | 0.9511 | 0.0000 | 0.1923 |
| | 16 | RAN | 0.8160 | 0.9187 | 0.7108 | 0.0618 | 0.6036 | 0.9698 | 0.0000 | 0.2591 |
| | | SAL | 0.8231 | 0.9158 | 0.7345 | 0.0406 | 0.6039 | 0.9691 | 0.0000 | 0.2614 |
| | | STL | 0.6473 | 0.8219 | 0.3878 | 0.1176 | 0.6273 | 0.9485 | 0.0000 | 0.2013 |

Table E.2: grouped-liblinear-enron-GM1-ALL-ALL-10

| GM1 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | RAN | 0.7373 | 0.8073 | 0.6803 | 0.0411 | 0.5176 | 0.9640 | 0.0000 | 0.2682 |
| | | SAL | 0.7414 | 0.7815 | 0.7195 | 0.0219 | 0.5144 | 0.9546 | 0.0000 | 0.2664 |
| | | STL | 0.5747 | 0.7400 | 0.4430 | 0.1079 | 0.5435 | 0.9337 | 0.0000 | 0.2273 |
| | 1 | RAN | 0.7300 | 0.8064 | 0.6522 | 0.0486 | 0.5081 | 0.9572 | 0.0000 | 0.2699 |
| | | SAL | 0.7511 | 0.7905 | 0.7100 | 0.0280 | 0.5241 | 0.9551 | 0.0000 | 0.2654 |
| | | STL | 0.5730 | 0.7486 | 0.4574 | 0.1109 | 0.5376 | 0.9347 | 0.0000 | 0.2323 |
| | 2 | RAN | 0.7442 | 0.8364 | 0.6904 | 0.0527 | 0.5211 | 0.9606 | 0.0000 | 0.2637 |
| | | SAL | 0.7450 | 0.7880 | 0.7076 | 0.0280 | 0.5141 | 0.9645 | 0.0000 | 0.2666 |
| | | STL | 0.5728 | 0.7472 | 0.4500 | 0.1120 | 0.5415 | 0.9427 | 0.0000 | 0.2285 |
| | 4 | RAN | 0.7306 | 0.7668 | 0.6556 | 0.0452 | 0.5175 | 0.9558 | 0.0000 | 0.2668 |
| | | SAL | 0.7454 | 0.7836 | 0.7137 | 0.0219 | 0.5147 | 0.9496 | 0.0000 | 0.2689 |
| | | STL | 0.5696 | 0.7409 | 0.4483 | 0.1095 | 0.5390 | 0.9388 | 0.0000 | 0.2277 |
| | 8 | RAN | 0.7383 | 0.8013 | 0.6680 | 0.0407 | 0.5254 | 0.9566 | 0.0000 | 0.2590 |
| | | SAL | 0.7476 | 0.7933 | 0.7244 | 0.0245 | 0.5183 | 0.9565 | 0.0000 | 0.2656 |
| | | STL | 0.5727 | 0.7454 | 0.4599 | 0.1051 | 0.5437 | 0.9441 | 0.0000 | 0.2264 |
| | 16 | RAN | 0.7364 | 0.7925 | 0.6755 | 0.0386 | 0.5113 | 0.9525 | 0.0000 | 0.2764 |
| | | SAL | 0.7455 | 0.7776 | 0.7148 | 0.0232 | 0.5132 | 0.9568 | 0.0000 | 0.2719 |
| | | STL | 0.5764 | 0.7487 | 0.4483 | 0.1090 | 0.5424 | 0.9388 | 0.0000 | 0.2344 |

Table E.3: grouped-liblinear-enron-GM1-ALL-ALL-25

| GM1 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | RAN | 0.6872 | 0.7476 | 0.6382 | 0.0454 | 0.4701 | 0.9509 | 0.0000 | 0.2677 |
| | | SAL | 0.6889 | 0.7028 | 0.6648 | 0.0171 | 0.4733 | 0.9426 | 0.0000 | 0.2611 |
| | | STL | 0.5472 | 0.6660 | 0.4341 | 0.0948 | 0.4966 | 0.9455 | 0.0000 | 0.2368 |
| | 1 | RAN | 0.6765 | 0.7280 | 0.5821 | 0.0669 | 0.4680 | 0.9486 | 0.0000 | 0.2701 |
| | | SAL | 0.6918 | 0.7114 | 0.6621 | 0.0214 | 0.4689 | 0.9561 | 0.0000 | 0.2625 |
| | | STL | 0.5408 | 0.6699 | 0.4234 | 0.1010 | 0.4870 | 0.9423 | 0.0000 | 0.2372 |
| | 2 | RAN | 0.6981 | 0.7214 | 0.6835 | 0.0167 | 0.4779 | 0.9475 | 0.0000 | 0.2691 |
| | | SAL | 0.6834 | 0.6975 | 0.6662 | 0.0130 | 0.4600 | 0.9426 | 0.0000 | 0.2633 |
| | | STL | 0.5443 | 0.6732 | 0.4287 | 0.1003 | 0.4912 | 0.9330 | 0.0000 | 0.2388 |
| | 4 | RAN | 0.6734 | 0.7052 | 0.6472 | 0.0240 | 0.4622 | 0.9395 | 0.0000 | 0.2671 |
| | | SAL | 0.6928 | 0.7026 | 0.6810 | 0.0089 | 0.4727 | 0.9532 | 0.0000 | 0.2677 |
| | | STL | 0.5407 | 0.6664 | 0.4327 | 0.0962 | 0.4904 | 0.9396 | 0.0000 | 0.2365 |
| | 8 | RAN | 0.6889 | 0.7127 | 0.6465 | 0.0301 | 0.4727 | 0.9511 | 0.0000 | 0.2694 |
| | | SAL | 0.6921 | 0.7013 | 0.6753 | 0.0119 | 0.4700 | 0.9559 | 0.0000 | 0.2724 |
| | | STL | 0.5447 | 0.6779 | 0.4376 | 0.0998 | 0.4911 | 0.9341 | 0.0000 | 0.2389 |
| | 16 | RAN | 0.6861 | 0.7524 | 0.6315 | 0.0500 | 0.4633 | 0.9416 | 0.0000 | 0.2699 |
| | | SAL | 0.6920 | 0.7105 | 0.6715 | 0.0160 | 0.4732 | 0.9504 | 0.0000 | 0.2648 |
| | | STL | 0.5530 | 0.6826 | 0.4406 | 0.0996 | 0.4948 | 0.9381 | 0.0000 | 0.2358 |

Table E.4: grouped-liblinear-enron-GM1-ALL-ALL-50

| GM1 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | Group Type | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | RAN | 0.6490 | 0.6499 | 0.6482 | 0.0009 | 0.4426 | 0.9437 | 0.0000 | 0.2615 |
| | | SAL | 0.6631 | 0.7024 | 0.6239 | 0.0393 | 0.4462 | 0.9388 | 0.0000 | 0.2682 |
| | | STL | 0.5316 | 0.6478 | 0.4155 | 0.1162 | 0.4589 | 0.9217 | 0.0000 | 0.2459 |
| | 1 | RAN | 0.6585 | 0.6724 | 0.6446 | 0.0139 | 0.4423 | 0.9365 | 0.0000 | 0.2678 |
| | | SAL | 0.6565 | 0.7011 | 0.6118 | 0.0447 | 0.4354 | 0.9394 | 0.0000 | 0.2714 |
| | | STL | 0.5155 | 0.6430 | 0.3880 | 0.1275 | 0.4445 | 0.9453 | 0.0000 | 0.2500 |
| | 2 | RAN | 0.6511 | 0.6786 | 0.6237 | 0.0275 | 0.4416 | 0.9511 | 0.0000 | 0.2647 |
| | | SAL | 0.6618 | 0.6858 | 0.6378 | 0.0240 | 0.4435 | 0.9286 | 0.0000 | 0.2644 |
| | | STL | 0.5253 | 0.6511 | 0.3995 | 0.1258 | 0.4534 | 0.9104 | 0.0000 | 0.2454 |
| | 4 | RAN | 0.6570 | 0.6793 | 0.6347 | 0.0223 | 0.4415 | 0.9402 | 0.0000 | 0.2681 |
| | | SAL | 0.6581 | 0.6814 | 0.6347 | 0.0234 | 0.4337 | 0.9281 | 0.0000 | 0.2662 |
| | | STL | 0.5245 | 0.6484 | 0.4006 | 0.1239 | 0.4501 | 0.9380 | 0.0000 | 0.2476 |
| | 8 | RAN | 0.6445 | 0.6632 | 0.6259 | 0.0186 | 0.4324 | 0.9392 | 0.0000 | 0.2626 |
| | | SAL | 0.6585 | 0.6836 | 0.6334 | 0.0251 | 0.4495 | 0.9275 | 0.0000 | 0.2627 |
| | | STL | 0.5226 | 0.6378 | 0.4074 | 0.1152 | 0.4479 | 0.9339 | 0.0000 | 0.2462 |
| | 16 | RAN | 0.6553 | 0.6716 | 0.6391 | 0.0163 | 0.4339 | 0.9413 | 0.0000 | 0.2676 |
| | | SAL | 0.6593 | 0.6655 | 0.6531 | 0.0062 | 0.4362 | 0.9401 | 0.0000 | 0.2647 |
| | | STL | 0.5264 | 0.6497 | 0.4030 | 0.1233 | 0.4507 | 0.9156 | 0.0000 | 0.2452 |

Table E.5: grouped-liblinear-enron-GM1-ALL-ALL-75

| GM1 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | RAN | 0.6033 | 0.6033 | 0.6033 | 0.0000 | 0.3951 | 0.9144 | 0.0000 | 0.2613 |
| | | SAL | 0.6074 | 0.6074 | 0.6074 | 0.0000 | 0.4025 | 0.9316 | 0.0000 | 0.2607 |
| | | STL | 0.6074 | 0.6074 | 0.6074 | 0.0000 | 0.4025 | 0.9316 | 0.0000 | 0.2607 |
| | 1 | RAN | 0.6155 | 0.6155 | 0.6155 | 0.0000 | 0.4059 | 0.9176 | 0.0000 | 0.2649 |
| | | SAL | 0.5849 | 0.5849 | 0.5849 | 0.0000 | 0.3923 | 0.9402 | 0.0000 | 0.2691 |
| | | STL | 0.5849 | 0.5849 | 0.5849 | 0.0000 | 0.3923 | 0.9402 | 0.0000 | 0.2691 |
| | 2 | RAN | 0.6049 | 0.6049 | 0.6049 | 0.0000 | 0.3968 | 0.9342 | 0.0000 | 0.2654 |
| | | SAL | 0.5949 | 0.5949 | 0.5949 | 0.0000 | 0.3954 | 0.9389 | 0.0000 | 0.2686 |
| | | STL | 0.5949 | 0.5949 | 0.5949 | 0.0000 | 0.3954 | 0.9389 | 0.0000 | 0.2686 |
| | 4 | RAN | 0.6065 | 0.6065 | 0.6065 | 0.0000 | 0.4026 | 0.9132 | 0.0000 | 0.2608 |
| | | SAL | 0.6093 | 0.6093 | 0.6093 | 0.0000 | 0.4042 | 0.9488 | 0.0000 | 0.2656 |
| | | STL | 0.6093 | 0.6093 | 0.6093 | 0.0000 | 0.4042 | 0.9488 | 0.0000 | 0.2656 |
| | 8 | RAN | 0.6008 | 0.6008 | 0.6008 | 0.0000 | 0.4015 | 0.9394 | 0.0000 | 0.2651 |
| | | SAL | 0.5982 | 0.5982 | 0.5982 | 0.0000 | 0.4028 | 0.9451 | 0.0000 | 0.2632 |
| | | STL | 0.5982 | 0.5982 | 0.5982 | 0.0000 | 0.4028 | 0.9451 | 0.0000 | 0.2632 |
| | 16 | RAN | 0.6011 | 0.6011 | 0.6011 | 0.0000 | 0.4038 | 0.9274 | 0.0000 | 0.2614 |
| | | SAL | 0.5975 | 0.5975 | 0.5975 | 0.0000 | 0.3968 | 0.9489 | 0.0000 | 0.2689 |
| | | STL | 0.5975 | 0.5975 | 0.5975 | 0.0000 | 0.3968 | 0.9489 | 0.0000 | 0.2689 |

Table E.6: grouped-liblinear-enron-GM1-ALL-ALL-150

| GM2 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | Group Type | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | RAN | 0.8948 | 0.9753 | 0.7918 | 0.0485 | 0.7299 | 1.0000 | 0.0000 | 0.2440 |
| | | SAL | 0.9091 | 0.9599 | 0.8276 | 0.0336 | 0.7160 | 0.9901 | 0.0000 | 0.2660 |
| | | STL | 0.7782 | 0.9509 | 0.5185 | 0.1226 | 0.7467 | 0.9847 | 0.0000 | 0.2057 |
| | 1 | RAN | 0.8538 | 0.9413 | 0.6271 | 0.0733 | 0.6626 | 0.9781 | 0.0000 | 0.2548 |
| | | SAL | 0.8745 | 0.9544 | 0.7669 | 0.0413 | 0.6615 | 0.9780 | 0.0000 | 0.2644 |
| | | STL | 0.7294 | 0.9356 | 0.4444 | 0.1127 | 0.7042 | 0.9676 | 0.0000 | 0.1876 |
| | 2 | RAN | 0.8520 | 0.9448 | 0.7456 | 0.0570 | 0.6654 | 0.9771 | 0.0000 | 0.2605 |
| | | SAL | 0.8758 | 0.9369 | 0.7792 | 0.0411 | 0.6651 | 0.9778 | 0.0000 | 0.2601 |
| | | STL | 0.7300 | 0.9133 | 0.4444 | 0.1143 | 0.7041 | 0.9658 | 0.0000 | 0.1903 |
| | 4 | RAN | 0.8487 | 0.9560 | 0.7092 | 0.0660 | 0.6539 | 0.9834 | 0.0000 | 0.2618 |
| | | SAL | 0.8775 | 0.9366 | 0.7835 | 0.0366 | 0.6662 | 0.9789 | 0.0000 | 0.2620 |
| | | STL | 0.7298 | 0.9381 | 0.4444 | 0.1136 | 0.7041 | 0.9672 | 0.0000 | 0.1896 |
| | 8 | RAN | 0.8532 | 0.9547 | 0.7217 | 0.0652 | 0.6544 | 0.9817 | 0.0000 | 0.2670 |
| | | SAL | 0.8783 | 0.9366 | 0.7688 | 0.0411 | 0.6671 | 0.9781 | 0.0000 | 0.2608 |
| | | STL | 0.7281 | 0.9262 | 0.4444 | 0.1120 | 0.7028 | 0.9609 | 0.0000 | 0.1869 |
| | 16 | RAN | 0.8440 | 0.9606 | 0.6183 | 0.0813 | 0.6695 | 0.9881 | 0.0000 | 0.2538 |
| | | SAL | 0.8733 | 0.9428 | 0.7773 | 0.0434 | 0.6605 | 0.9789 | 0.0000 | 0.2653 |
| | | STL | 0.7291 | 0.9077 | 0.4444 | 0.1101 | 0.7047 | 0.9680 | 0.0000 | 0.1842 |

Table E.7: grouped-liblinear-enron-GM2-ALL-ALL-5

| GM2 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | RAN | 0.8552 | 0.9369 | 0.7431 | 0.0453 | 0.6870 | 0.9789 | 0.0000 | 0.2520 |
| | | SAL | 0.8708 | 0.9297 | 0.8023 | 0.0350 | 0.6791 | 0.9811 | 0.0000 | 0.2612 |
| | | STL | 0.7188 | 0.9258 | 0.5000 | 0.1243 | 0.6945 | 0.9802 | 0.0000 | 0.2174 |
| | 1 | RAN | 0.8038 | 0.9093 | 0.6536 | 0.0624 | 0.5769 | 0.9711 | 0.0000 | 0.2683 |
| | | SAL | 0.8177 | 0.8707 | 0.7580 | 0.0376 | 0.5879 | 0.9616 | 0.0000 | 0.2696 |
| | | STL | 0.6438 | 0.8086 | 0.3936 | 0.1151 | 0.6178 | 0.9592 | 0.0000 | 0.2084 |
| | 2 | RAN | 0.8181 | 0.9255 | 0.7252 | 0.0584 | 0.5925 | 0.9711 | 0.0000 | 0.2721 |
| | | SAL | 0.8107 | 0.8651 | 0.7298 | 0.0394 | 0.5878 | 0.9641 | 0.0000 | 0.2621 |
| | | STL | 0.6447 | 0.8129 | 0.3936 | 0.1184 | 0.6199 | 0.9537 | 0.0000 | 0.2138 |
| | 4 | RAN | 0.7917 | 0.8918 | 0.6272 | 0.0747 | 0.5913 | 0.9675 | 0.0000 | 0.2618 |
| | | SAL | 0.8148 | 0.8969 | 0.7263 | 0.0476 | 0.5891 | 0.9622 | 0.0000 | 0.2704 |
| | | STL | 0.6440 | 0.8130 | 0.3936 | 0.1141 | 0.6191 | 0.9535 | 0.0000 | 0.2057 |
| | 8 | RAN | 0.8112 | 0.8739 | 0.6976 | 0.0508 | 0.5855 | 0.9634 | 0.0000 | 0.2628 |
| | | SAL | 0.8194 | 0.8722 | 0.7347 | 0.0398 | 0.5958 | 0.9678 | 0.0000 | 0.2667 |
| | | STL | 0.6437 | 0.8110 | 0.3936 | 0.1181 | 0.6195 | 0.9561 | 0.0000 | 0.2120 |
| | 16 | RAN | 0.8008 | 0.9065 | 0.6550 | 0.0763 | 0.5866 | 0.9728 | 0.0000 | 0.2670 |
| | | SAL | 0.8128 | 0.8887 | 0.6917 | 0.0460 | 0.5886 | 0.9615 | 0.0000 | 0.2704 |
| | | STL | 0.6448 | 0.8073 | 0.3936 | 0.1155 | 0.6205 | 0.9596 | 0.0000 | 0.2093 |

Table E.8: grouped-liblinear-enron-GM2-ALL-ALL-10

| GM2 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | RAN | 0.8191 | 0.8989 | 0.7154 | 0.0639 | 0.6266 | 0.9810 | 0.0000 | 0.2710 |
| | | SAL | 0.8224 | 0.8598 | 0.7468 | 0.0363 | 0.6270 | 0.9814 | 0.0000 | 0.2725 |
| | | STL | 0.6674 | 0.8551 | 0.4824 | 0.1262 | 0.6416 | 0.9748 | 0.0000 | 0.2442 |
| | 1 | RAN | 0.7356 | 0.7693 | 0.6908 | 0.0296 | 0.5124 | 0.9610 | 0.0000 | 0.2729 |
| | | SAL | 0.7309 | 0.8075 | 0.6765 | 0.0392 | 0.5039 | 0.9632 | 0.0000 | 0.2700 |
| | | STL | 0.5704 | 0.7427 | 0.4512 | 0.1123 | 0.5351 | 0.9518 | 0.0000 | 0.2313 |
| | 2 | RAN | 0.7394 | 0.8065 | 0.6931 | 0.0411 | 0.5174 | 0.9526 | 0.0000 | 0.2680 |
| | | SAL | 0.7353 | 0.8036 | 0.6822 | 0.0411 | 0.5119 | 0.9612 | 0.0000 | 0.2695 |
| | | STL | 0.5718 | 0.7216 | 0.4562 | 0.0986 | 0.5343 | 0.9441 | 0.0000 | 0.2270 |
| | 4 | RAN | 0.7403 | 0.7758 | 0.6806 | 0.0344 | 0.5097 | 0.9512 | 0.0000 | 0.2696 |
| | | SAL | 0.7324 | 0.7994 | 0.7053 | 0.0327 | 0.5094 | 0.9674 | 0.0000 | 0.2643 |
| | | STL | 0.5701 | 0.7473 | 0.4465 | 0.1126 | 0.5329 | 0.9510 | 0.0000 | 0.2334 |
| | 8 | RAN | 0.7192 | 0.8331 | 0.6238 | 0.0661 | 0.5093 | 0.9655 | 0.0000 | 0.2689 |
| | | SAL | 0.7362 | 0.8000 | 0.6915 | 0.0359 | 0.5106 | 0.9562 | 0.0000 | 0.2677 |
| | | STL | 0.5751 | 0.7448 | 0.4486 | 0.1055 | 0.5390 | 0.9492 | 0.0000 | 0.2269 |
| | 16 | RAN | 0.7317 | 0.8142 | 0.6307 | 0.0670 | 0.5174 | 0.9613 | 0.0000 | 0.2645 |
| | | SAL | 0.7302 | 0.8380 | 0.6722 | 0.0531 | 0.5027 | 0.9628 | 0.0000 | 0.2685 |
| | | STL | 0.5661 | 0.7312 | 0.4470 | 0.1040 | 0.5286 | 0.9518 | 0.0000 | 0.2300 |

Table E.9: grouped-liblinear-enron-GM2-ALL-ALL-25

| GM2 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | RAN | 0.7702 | 0.8240 | 0.6807 | 0.0637 | 0.5973 | 0.9747 | 0.0000 | 0.2689 |
| | | SAL | 0.7940 | 0.8240 | 0.7423 | 0.0367 | 0.5922 | 0.9780 | 0.0000 | 0.2791 |
| | | STL | 0.6564 | 0.8029 | 0.5252 | 0.1139 | 0.6109 | 0.9731 | 0.0000 | 0.2545 |
| | 1 | RAN | 0.6743 | 0.7015 | 0.6543 | 0.0199 | 0.4607 | 0.9505 | 0.0000 | 0.2681 |
| | | SAL | 0.6892 | 0.7216 | 0.6673 | 0.0234 | 0.4598 | 0.9480 | 0.0000 | 0.2718 |
| | | STL | 0.5354 | 0.6679 | 0.4211 | 0.1016 | 0.4843 | 0.9348 | 0.0000 | 0.2432 |
| | 2 | RAN | 0.6817 | 0.7054 | 0.6572 | 0.0197 | 0.4625 | 0.9318 | 0.0000 | 0.2650 |
| | | SAL | 0.6890 | 0.7156 | 0.6716 | 0.0191 | 0.4677 | 0.9504 | 0.0000 | 0.2658 |
| | | STL | 0.5317 | 0.6701 | 0.4119 | 0.1062 | 0.4826 | 0.9357 | 0.0000 | 0.2425 |
| | 4 | RAN | 0.6806 | 0.7174 | 0.6552 | 0.0266 | 0.4638 | 0.9460 | 0.0000 | 0.2712 |
| | | SAL | 0.6808 | 0.7030 | 0.6581 | 0.0183 | 0.4587 | 0.9540 | 0.0000 | 0.2726 |
| | | STL | 0.5343 | 0.6607 | 0.4211 | 0.0983 | 0.4837 | 0.9347 | 0.0000 | 0.2387 |
| | 8 | RAN | 0.6771 | 0.7296 | 0.6471 | 0.0372 | 0.4545 | 0.9333 | 0.0000 | 0.2730 |
| | | SAL | 0.6883 | 0.7290 | 0.6577 | 0.0300 | 0.4729 | 0.9402 | 0.0000 | 0.2669 |
| | | STL | 0.5325 | 0.6714 | 0.4092 | 0.1076 | 0.4786 | 0.9327 | 0.0000 | 0.2393 |

Table E.10: grouped-liblinear-enron-GM2-ALL-ALL-50

| GM2 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | RAN | 0.7667 | 0.7969 | 0.7365 | 0.0302 | 0.5812 | 0.9747 | 0.0000 | 0.2717 |
| | | SAL | 0.7747 | 0.7839 | 0.7656 | 0.0092 | 0.5797 | 0.9786 | 0.0000 | 0.2763 |
| | | STL | 0.6576 | 0.7716 | 0.5437 | 0.1139 | 0.5887 | 0.9737 | 0.0000 | 0.2682 |
| | 1 | RAN | 0.6565 | 0.6959 | 0.6171 | 0.0394 | 0.4408 | 0.9313 | 0.0000 | 0.2716 |
| | | SAL | 0.6640 | 0.6751 | 0.6528 | 0.0112 | 0.4415 | 0.9564 | 0.0000 | 0.2767 |
| | | STL | 0.5156 | 0.6440 | 0.3872 | 0.1284 | 0.4539 | 0.9444 | 0.0000 | 0.2453 |
| | 2 | RAN | 0.6439 | 0.6625 | 0.6253 | 0.0186 | 0.4317 | 0.9523 | 0.0000 | 0.2687 |
| | | SAL | 0.6329 | 0.6677 | 0.5980 | 0.0349 | 0.4301 | 0.9428 | 0.0000 | 0.2722 |
| | | STL | 0.5194 | 0.6514 | 0.3874 | 0.1320 | 0.4538 | 0.9377 | 0.0000 | 0.2538 |
| | 4 | RAN | 0.6397 | 0.6441 | 0.6352 | 0.0045 | 0.4354 | 0.9450 | 0.0000 | 0.2725 |
| | | SAL | 0.6565 | 0.6913 | 0.6216 | 0.0349 | 0.4427 | 0.9492 | 0.0000 | 0.2666 |
| | | STL | 0.5209 | 0.6467 | 0.3950 | 0.1259 | 0.4609 | 0.9304 | 0.0000 | 0.2508 |
| | 8 | RAN | 0.6570 | 0.6647 | 0.6492 | 0.0077 | 0.4352 | 0.9490 | 0.0000 | 0.2731 |
| | | SAL | 0.6420 | 0.6681 | 0.6158 | 0.0262 | 0.4304 | 0.9344 | 0.0000 | 0.2696 |
| | | STL | 0.5112 | 0.6471 | 0.3752 | 0.1359 | 0.4470 | 0.9428 | 0.0000 | 0.2535 |

Table E.11: grouped-liblinear-enron-GM2-ALL-ALL-75

| GM2 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | Group Type | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | RAN | 0.7429 | 0.7429 | 0.7429 | 0.0000 | 0.5493 | 0.9731 | 0.0000 | 0.2777 |
| | | SAL | 0.7456 | 0.7456 | 0.7456 | 0.0000 | 0.5528 | 0.9737 | 0.0000 | 0.2798 |
| | | STL | 0.7456 | 0.7456 | 0.7456 | 0.0000 | 0.5528 | 0.9737 | 0.0000 | 0.2798 |
| | 1 | RAN | 0.5841 | 0.5841 | 0.5841 | 0.0000 | 0.3870 | 0.9285 | 0.0000 | 0.2690 |
| | | SAL | 0.6047 | 0.6047 | 0.6047 | 0.0000 | 0.4034 | 0.9387 | 0.0000 | 0.2698 |
| | | STL | 0.6047 | 0.6047 | 0.6047 | 0.0000 | 0.4034 | 0.9387 | 0.0000 | 0.2698 |
| | 2 | RAN | 0.6134 | 0.6134 | 0.6134 | 0.0000 | 0.4057 | 0.9311 | 0.0000 | 0.2653 |
| | | SAL | 0.6105 | 0.6105 | 0.6105 | 0.0000 | 0.4030 | 0.9283 | 0.0000 | 0.2668 |
| | | STL | 0.6105 | 0.6105 | 0.6105 | 0.0000 | 0.4030 | 0.9283 | 0.0000 | 0.2668 |
| | 4 | RAN | 0.6057 | 0.6057 | 0.6057 | 0.0000 | 0.4016 | 0.9327 | 0.0000 | 0.2717 |
| | | SAL | 0.6073 | 0.6073 | 0.6073 | 0.0000 | 0.4010 | 0.9391 | 0.0000 | 0.2709 |
| | | STL | 0.6073 | 0.6073 | 0.6073 | 0.0000 | 0.4010 | 0.9391 | 0.0000 | 0.2709 |

Table E.12: grouped-liblinear-enron-GM2-ALL-ALL-150

| GM5 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | Group Type | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | RAN | 0.7207 | 0.9636 | 0.4010 | 0.1429 | 0.5052 | 1.0000 | 0.0000 | 0.3099 |
| | | SAL | 0.7744 | 0.9271 | 0.5114 | 0.1158 | 0.5026 | 1.0000 | 0.0000 | 0.3307 |
| | | STL | 0.5692 | 0.8742 | 0.3017 | 0.1344 | 0.5108 | 1.0000 | 0.0000 | 0.2582 |
| | 1 | RAN | 0.8487 | 0.9685 | 0.7000 | 0.0784 | 0.6751 | 0.9869 | 0.0000 | 0.2565 |
| | | SAL | 0.8705 | 0.9378 | 0.7535 | 0.0455 | 0.6656 | 0.9824 | 0.0000 | 0.2587 |
| | | STL | 0.7160 | 0.9274 | 0.4000 | 0.1367 | 0.6913 | 0.9590 | 0.0000 | 0.2018 |
| | 2 | RAN | 0.8515 | 0.9550 | 0.5736 | 0.0716 | 0.6635 | 0.9836 | 0.0000 | 0.2600 |
| | | SAL | 0.8636 | 0.9498 | 0.7201 | 0.0609 | 0.6591 | 0.9824 | 0.0000 | 0.2612 |
| | | STL | 0.7201 | 0.9287 | 0.4000 | 0.1304 | 0.6949 | 0.9613 | 0.0000 | 0.1986 |
| | 4 | RAN | 0.8531 | 0.9421 | 0.6677 | 0.0616 | 0.6730 | 0.9714 | 0.0000 | 0.2549 |
| | | SAL | 0.8696 | 0.9455 | 0.7224 | 0.0535 | 0.6671 | 0.9821 | 0.0000 | 0.2582 |
| | | STL | 0.7163 | 0.9264 | 0.4000 | 0.1352 | 0.6915 | 0.9666 | 0.0000 | 0.2014 |
| | 8 | RAN | 0.8383 | 0.9513 | 0.6020 | 0.0744 | 0.6728 | 0.9700 | 0.0000 | 0.2452 |
| | | SAL | 0.8705 | 0.9470 | 0.7557 | 0.0454 | 0.6656 | 0.9824 | 0.0000 | 0.2597 |
| | | STL | 0.7122 | 0.9284 | 0.4000 | 0.1349 | 0.6864 | 0.9639 | 0.0000 | 0.2006 |
| | 16 | RAN | 0.8522 | 0.9676 | 0.6954 | 0.0732 | 0.6686 | 0.9874 | 0.0000 | 0.2547 |
| | | SAL | 0.8676 | 0.9463 | 0.7557 | 0.0519 | 0.6630 | 0.9824 | 0.0000 | 0.2603 |
| | | STL | 0.7138 | 0.9276 | 0.4000 | 0.1365 | 0.6881 | 0.9630 | 0.0000 | 0.2014 |

Table E.13: grouped-liblinear-enron-GM5-ALL-ALL-5

| GM5 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | RAN | 0.6930 | 0.8560 | 0.4644 | 0.0931 | 0.4657 | 0.9867 | 0.0000 | 0.3066 |
| | | SAL | 0.7104 | 0.8020 | 0.5628 | 0.0630 | 0.4652 | 0.9870 | 0.0000 | 0.3143 |
| | | STL | 0.4857 | 0.7801 | 0.2548 | 0.1328 | 0.4507 | 0.9870 | 0.0000 | 0.2766 |
| | 1 | RAN | 0.8095 | 0.9022 | 0.7286 | 0.0564 | 0.5826 | 0.9782 | 0.0000 | 0.2630 |
| | | SAL | 0.8057 | 0.8903 | 0.6153 | 0.0603 | 0.5886 | 0.9727 | 0.0000 | 0.2639 |
| | | STL | 0.6406 | 0.8044 | 0.3908 | 0.1168 | 0.6225 | 0.9604 | 0.0000 | 0.2110 |
| | 2 | RAN | 0.7999 | 0.9145 | 0.7189 | 0.0572 | 0.5902 | 0.9684 | 0.0000 | 0.2723 |
| | | SAL | 0.7957 | 0.9221 | 0.6462 | 0.0685 | 0.5860 | 0.9733 | 0.0000 | 0.2568 |
| | | STL | 0.6365 | 0.7796 | 0.3908 | 0.1135 | 0.6181 | 0.9529 | 0.0000 | 0.2119 |
| | 4 | RAN | 0.7865 | 0.9187 | 0.5834 | 0.0723 | 0.5908 | 0.9714 | 0.0000 | 0.2641 |
| | | SAL | 0.7985 | 0.9256 | 0.5951 | 0.0810 | 0.5848 | 0.9729 | 0.0000 | 0.2634 |
| | | STL | 0.6405 | 0.7866 | 0.3908 | 0.1150 | 0.6220 | 0.9551 | 0.0000 | 0.2108 |
| | 8 | RAN | 0.8017 | 0.9272 | 0.6714 | 0.0691 | 0.5895 | 0.9764 | 0.0000 | 0.2584 |
| | | SAL | 0.8020 | 0.9232 | 0.5924 | 0.0707 | 0.5874 | 0.9736 | 0.0000 | 0.2637 |
| | | STL | 0.6394 | 0.8052 | 0.3908 | 0.1191 | 0.6214 | 0.9565 | 0.0000 | 0.2104 |
| | 16 | RAN | 0.8086 | 0.9142 | 0.6593 | 0.0748 | 0.6006 | 0.9770 | 0.0000 | 0.2613 |
| | | SAL | 0.8080 | 0.9191 | 0.6835 | 0.0576 | 0.5901 | 0.9734 | 0.0000 | 0.2652 |
| | | STL | 0.6422 | 0.8006 | 0.3908 | 0.1194 | 0.6246 | 0.9573 | 0.0000 | 0.2122 |

Table E.14: grouped-liblinear-enron-GM5-ALL-ALL-10

| GM5 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | Group Type | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | RAN | 0.5816 | 0.7076 | 0.4018 | 0.1110 | 0.4370 | 1.0000 | 0.0000 | 0.3091 |
| | | SAL | 0.6305 | 0.6898 | 0.5050 | 0.0599 | 0.4377 | 0.9744 | 0.0000 | 0.3124 |
| | | STL | 0.4377 | 0.6993 | 0.3305 | 0.1268 | 0.4368 | 0.9870 | 0.0000 | 0.3003 |
| | 1 | RAN | 0.7379 | 0.7867 | 0.6328 | 0.0574 | 0.5186 | 0.9523 | 0.0000 | 0.2697 |
| | | SAL | 0.7247 | 0.7775 | 0.6392 | 0.0455 | 0.5128 | 0.9554 | 0.0000 | 0.2673 |
| | | STL | 0.5652 | 0.7111 | 0.4371 | 0.1041 | 0.5413 | 0.9412 | 0.0000 | 0.2270 |
| | 2 | RAN | 0.7282 | 0.7445 | 0.7007 | 0.0137 | 0.5224 | 0.9551 | 0.0000 | 0.2684 |
| | | SAL | 0.7287 | 0.7759 | 0.6819 | 0.0279 | 0.5155 | 0.9599 | 0.0000 | 0.2666 |
| | | STL | 0.5616 | 0.7027 | 0.4371 | 0.0970 | 0.5378 | 0.9395 | 0.0000 | 0.2282 |
| | 4 | RAN | 0.7235 | 0.8209 | 0.6430 | 0.0658 | 0.5187 | 0.9488 | 0.0000 | 0.2720 |
| | | SAL | 0.7320 | 0.7808 | 0.6872 | 0.0288 | 0.5180 | 0.9629 | 0.0000 | 0.2654 |
| | | STL | 0.5607 | 0.7156 | 0.4371 | 0.1032 | 0.5388 | 0.9390 | 0.0000 | 0.2273 |

Table E.15: grouped-liblinear-enron-GM5-ALL-ALL-25

| GM5 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | Group Type | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | RAN | 0.5793 | 0.6564 | 0.4924 | 0.0673 | 0.4272 | 0.9870 | 0.0000 | 0.3154 |
| | | SAL | 0.5812 | 0.6174 | 0.5616 | 0.0256 | 0.4256 | 0.9870 | 0.0000 | 0.3083 |
| | | STL | 0.4365 | 0.5943 | 0.3088 | 0.1185 | 0.4259 | 0.9870 | 0.0000 | 0.3052 |
| | 1 | RAN | 0.6659 | 0.7342 | 0.5793 | 0.0646 | 0.4697 | 0.9572 | 0.0000 | 0.2773 |
| | | SAL | 0.6784 | 0.7134 | 0.6323 | 0.0340 | 0.4698 | 0.9571 | 0.0000 | 0.2709 |
| | | STL | 0.5333 | 0.6441 | 0.4304 | 0.0874 | 0.4877 | 0.9281 | 0.0000 | 0.2470 |
| | 2 | RAN | 0.6780 | 0.7022 | 0.6515 | 0.0208 | 0.4747 | 0.9584 | 0.0000 | 0.2702 |
| | | SAL | 0.6825 | 0.7004 | 0.6661 | 0.0140 | 0.4729 | 0.9610 | 0.0000 | 0.2749 |
| | | STL | 0.5452 | 0.6611 | 0.4442 | 0.0892 | 0.4976 | 0.9399 | 0.0000 | 0.2503 |

Table E.16: grouped-liblinear-enron-GM5-ALL-ALL-50

| GM5 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | Group Type | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | RAN | 0.5703 | 0.6206 | 0.5199 | 0.0503 | 0.4193 | 0.9870 | 0.0000 | 0.3119 |
| | | SAL | 0.5454 | 0.5772 | 0.5136 | 0.0318 | 0.4086 | 0.9870 | 0.0000 | 0.3215 |
| | | STL | 0.4475 | 0.5710 | 0.3240 | 0.1235 | 0.4164 | 0.9744 | 0.0000 | 0.3177 |
| | 1 | RAN | 0.6525 | 0.7194 | 0.5856 | 0.0669 | 0.4427 | 0.9513 | 0.0000 | 0.2758 |
| | | SAL | 0.6549 | 0.6679 | 0.6419 | 0.0130 | 0.4445 | 0.9497 | 0.0000 | 0.2763 |
| | | STL | 0.5350 | 0.6489 | 0.4211 | 0.1139 | 0.4672 | 0.9421 | 0.0000 | 0.2567 |
| | 2 | RAN | 0.6511 | 0.6739 | 0.6283 | 0.0228 | 0.4460 | 0.9437 | 0.0000 | 0.2699 |
| | | SAL | 0.6461 | 0.6551 | 0.6371 | 0.0090 | 0.4410 | 0.9494 | 0.0000 | 0.2737 |
| | | STL | 0.5341 | 0.6451 | 0.4230 | 0.1111 | 0.4674 | 0.9380 | 0.0000 | 0.2542 |

Table E.17: grouped-liblinear-enron-GM5-ALL-ALL-75

| GM5 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | Group Type | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | RAN | 0.5418 | 0.5418 | 0.5418 | 0.0000 | 0.4035 | 0.9737 | 0.0000 | 0.3182 |
| | | SAL | 0.5478 | 0.5478 | 0.5478 | 0.0000 | 0.4034 | 0.9870 | 0.0000 | 0.3169 |
| | | STL | 0.5478 | 0.5478 | 0.5478 | 0.0000 | 0.4034 | 0.9870 | 0.0000 | 0.3169 |
| | 1 | RAN | 0.6091 | 0.6091 | 0.6091 | 0.0000 | 0.4141 | 0.9493 | 0.0000 | 0.2691 |
| | | SAL | 0.5875 | 0.5875 | 0.5875 | 0.0000 | 0.4043 | 0.9502 | 0.0000 | 0.2656 |
| | | STL | 0.5875 | 0.5875 | 0.5875 | 0.0000 | 0.4043 | 0.9502 | 0.0000 | 0.2656 |

Table E.18: grouped-liblinear-enron-GM5-ALL-ALL-150

| GB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | RAN | 0.8782 | 0.9835 | 0.6177 | 0.0747 | 0.7112 | 1.0000 | 0.0000 | 0.2621 |
| | | SAL | 0.9047 | 0.9706 | 0.7487 | 0.0431 | 0.7088 | 1.0000 | 0.0000 | 0.2663 |
| | | STL | 0.7758 | 0.9533 | 0.5185 | 0.1190 | 0.7410 | 0.9839 | 0.0000 | 0.2063 |
| | 1 | RAN | 0.8781 | 0.9673 | 0.7660 | 0.0609 | 0.7147 | 0.9822 | 0.0000 | 0.2540 |
| | | SAL | 0.9020 | 0.9662 | 0.7589 | 0.0466 | 0.7005 | 0.9854 | 0.0000 | 0.2721 |
| | | STL | 0.7682 | 0.9664 | 0.5185 | 0.1219 | 0.7364 | 0.9837 | 0.0000 | 0.2054 |
| | 2 | RAN | 0.8740 | 0.9805 | 0.6998 | 0.0747 | 0.7114 | 0.9890 | 0.0000 | 0.2549 |
| | | SAL | 0.9014 | 0.9658 | 0.7589 | 0.0460 | 0.6988 | 0.9840 | 0.0000 | 0.2736 |
| | | STL | 0.7675 | 0.9632 | 0.5185 | 0.1215 | 0.7354 | 0.9834 | 0.0000 | 0.2061 |
| | 4 | RAN | 0.9046 | 0.9762 | 0.8386 | 0.0344 | 0.7172 | 0.9821 | 0.0000 | 0.2684 |
| | | SAL | 0.9016 | 0.9666 | 0.7585 | 0.0461 | 0.6994 | 0.9843 | 0.0000 | 0.2733 |
| | | STL | 0.7676 | 0.9663 | 0.5185 | 0.1220 | 0.7355 | 0.9844 | 0.0000 | 0.2062 |
| | 8 | RAN | 0.8921 | 0.9786 | 0.8128 | 0.0426 | 0.7116 | 0.9921 | 0.0000 | 0.2628 |
| | | SAL | 0.9025 | 0.9666 | 0.7581 | 0.0466 | 0.6999 | 0.9881 | 0.0000 | 0.2751 |
| | | STL | 0.7661 | 0.9641 | 0.5185 | 0.1210 | 0.7341 | 0.9826 | 0.0000 | 0.2057 |
| | 16 | RAN | 0.8881 | 0.9756 | 0.7216 | 0.0591 | 0.7070 | 0.9884 | 0.0000 | 0.2677 |
| | | SAL | 0.9025 | 0.9666 | 0.7605 | 0.0466 | 0.6996 | 0.9886 | 0.0000 | 0.2750 |
| | | STL | 0.7663 | 0.9643 | 0.5185 | 0.1218 | 0.7343 | 0.9821 | 0.0000 | 0.2063 |

Table E.19: grouped-liblinear-enron-GB3-ALL-ALL-5

| GB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | RAN | 0.8618 | 0.9407 | 0.7427 | 0.0577 | 0.6638 | 1.0000 | 0.0000 | 0.2742 |
| | | SAL | 0.8626 | 0.9538 | 0.7131 | 0.0564 | 0.6593 | 1.0000 | 0.0000 | 0.2753 |
| | | STL | 0.7128 | 0.9326 | 0.4787 | 0.1349 | 0.6867 | 0.9816 | 0.0000 | 0.2261 |
| | 1 | RAN | 0.8550 | 0.9341 | 0.7705 | 0.0454 | 0.6599 | 0.9818 | 0.0000 | 0.2701 |
| | | SAL | 0.8680 | 0.9242 | 0.6904 | 0.0569 | 0.6658 | 1.0000 | 0.0000 | 0.2701 |
| | | STL | 0.7151 | 0.9302 | 0.4894 | 0.1248 | 0.6880 | 0.9806 | 0.0000 | 0.2189 |
| | 2 | RAN | 0.8550 | 0.9265 | 0.7683 | 0.0454 | 0.6737 | 0.9819 | 0.0000 | 0.2588 |
| | | SAL | 0.8657 | 0.9221 | 0.6941 | 0.0562 | 0.6610 | 0.9870 | 0.0000 | 0.2715 |
| | | STL | 0.7176 | 0.9297 | 0.4894 | 0.1254 | 0.6912 | 0.9796 | 0.0000 | 0.2185 |
| | 4 | RAN | 0.8471 | 0.9426 | 0.7322 | 0.0584 | 0.6681 | 0.9870 | 0.0000 | 0.2658 |
| | | SAL | 0.8640 | 0.9204 | 0.7026 | 0.0546 | 0.6572 | 1.0000 | 0.0000 | 0.2751 |
| | | STL | 0.7177 | 0.9300 | 0.4894 | 0.1253 | 0.6915 | 0.9806 | 0.0000 | 0.2187 |
| | 8 | RAN | 0.8560 | 0.9512 | 0.7510 | 0.0590 | 0.6667 | 0.9808 | 0.0000 | 0.2654 |
| | | SAL | 0.8664 | 0.9185 | 0.6894 | 0.0577 | 0.6595 | 0.9870 | 0.0000 | 0.2748 |
| | | STL | 0.7177 | 0.9292 | 0.4894 | 0.1253 | 0.6912 | 0.9799 | 0.0000 | 0.2187 |
| | 16 | RAN | 0.8525 | 0.9215 | 0.7137 | 0.0541 | 0.6616 | 0.9831 | 0.0000 | 0.2682 |
| | | SAL | 0.8690 | 0.9264 | 0.7208 | 0.0513 | 0.6629 | 0.9870 | 0.0000 | 0.2722 |
| | | STL | 0.7170 | 0.9288 | 0.4894 | 0.1248 | 0.6906 | 0.9801 | 0.0000 | 0.2184 |

Table E.20: grouped-liblinear-enron-GB3-ALL-ALL-10

| GB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | RAN | 0.8190 | 0.8758 | 0.7683 | 0.0365 | 0.6151 | 0.9797 | 0.0000 | 0.2800 |
| | | SAL | 0.8195 | 0.8882 | 0.7559 | 0.0473 | 0.6142 | 0.9774 | 0.0000 | 0.2768 |
| | | STL | 0.6657 | 0.8529 | 0.5215 | 0.1221 | 0.6361 | 0.9700 | 0.0000 | 0.2437 |
| | 1 | RAN | 0.8158 | 0.8744 | 0.6640 | 0.0708 | 0.6150 | 0.9812 | 0.0000 | 0.2762 |
| | | SAL | 0.8235 | 0.8733 | 0.7936 | 0.0248 | 0.6158 | 0.9816 | 0.0000 | 0.2720 |
| | | STL | 0.6562 | 0.8501 | 0.5158 | 0.1235 | 0.6255 | 0.9717 | 0.0000 | 0.2480 |
| | 2 | RAN | 0.8268 | 0.8600 | 0.7862 | 0.0291 | 0.6207 | 0.9765 | 0.0000 | 0.2693 |
| | | SAL | 0.8258 | 0.8788 | 0.8011 | 0.0253 | 0.6174 | 0.9772 | 0.0000 | 0.2709 |
| | | STL | 0.6527 | 0.8497 | 0.5130 | 0.1264 | 0.6242 | 0.9721 | 0.0000 | 0.2533 |
| | 4 | RAN | 0.8230 | 0.9056 | 0.7260 | 0.0568 | 0.6185 | 0.9802 | 0.0000 | 0.2705 |
| | | SAL | 0.8238 | 0.8793 | 0.7973 | 0.0262 | 0.6152 | 0.9771 | 0.0000 | 0.2723 |
| | | STL | 0.6534 | 0.8498 | 0.5144 | 0.1258 | 0.6243 | 0.9721 | 0.0000 | 0.2526 |
| | 8 | RAN | 0.8088 | 0.8812 | 0.7701 | 0.0363 | 0.6109 | 0.9750 | 0.0000 | 0.2702 |
| | | SAL | 0.8230 | 0.8563 | 0.7995 | 0.0177 | 0.6137 | 0.9776 | 0.0000 | 0.2718 |
| | | STL | 0.6546 | 0.8535 | 0.5144 | 0.1255 | 0.6251 | 0.9720 | 0.0000 | 0.2507 |

Table E.21: grouped-liblinear-enron-GB3-ALL-ALL-25

| GB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | RAN | 0.7902 | 0.8149 | 0.7445 | 0.0323 | 0.5839 | 1.0000 | 0.0000 | 0.2793 |
| | | SAL | 0.7905 | 0.8465 | 0.7234 | 0.0509 | 0.5898 | 0.9761 | 0.0000 | 0.2784 |
| | | STL | 0.6421 | 0.8024 | 0.4980 | 0.1248 | 0.6004 | 0.9754 | 0.0000 | 0.2590 |
| | 1 | RAN | 0.7853 | 0.8094 | 0.7482 | 0.0266 | 0.5776 | 0.9753 | 0.0000 | 0.2820 |
| | | SAL | 0.7918 | 0.8204 | 0.7663 | 0.0222 | 0.5864 | 0.9709 | 0.0000 | 0.2792 |
| | | STL | 0.6344 | 0.8026 | 0.4731 | 0.1346 | 0.5955 | 0.9727 | 0.0000 | 0.2576 |
| | 2 | RAN | 0.7908 | 0.8058 | 0.7743 | 0.0129 | 0.5867 | 0.9753 | 0.0000 | 0.2803 |
| | | SAL | 0.7891 | 0.8100 | 0.7663 | 0.0179 | 0.5836 | 1.0000 | 0.0000 | 0.2756 |
| | | STL | 0.6326 | 0.8017 | 0.4765 | 0.1331 | 0.5960 | 0.9747 | 0.0000 | 0.2584 |
| | 4 | RAN | 0.7895 | 0.8628 | 0.7357 | 0.0537 | 0.5897 | 0.9759 | 0.0000 | 0.2759 |
| | | SAL | 0.7884 | 0.8137 | 0.7670 | 0.0193 | 0.5833 | 1.0000 | 0.0000 | 0.2790 |
| | | STL | 0.6305 | 0.7973 | 0.4735 | 0.1324 | 0.5938 | 0.9725 | 0.0000 | 0.2604 |

Table E.22: grouped-liblinear-enron-GB3-ALL-ALL-50

| GB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | RAN | 0.7731 | 0.7764 | 0.7698 | 0.0033 | 0.5722 | 0.9870 | 0.0000 | 0.2775 |
| | | SAL | 0.7662 | 0.7955 | 0.7368 | 0.0294 | 0.5603 | 0.9758 | 0.0000 | 0.2858 |
| | | STL | 0.6507 | 0.7794 | 0.5220 | 0.1287 | 0.5805 | 0.9744 | 0.0000 | 0.2711 |
| | 1 | RAN | 0.7762 | 0.8161 | 0.7364 | 0.0399 | 0.5810 | 0.9712 | 0.0000 | 0.2717 |
| | | SAL | 0.7740 | 0.7836 | 0.7644 | 0.0096 | 0.5735 | 0.9763 | 0.0000 | 0.2786 |
| | | STL | 0.6506 | 0.7739 | 0.5273 | 0.1233 | 0.5774 | 0.9718 | 0.0000 | 0.2662 |
| | 2 | RAN | 0.7734 | 0.7886 | 0.7583 | 0.0152 | 0.5686 | 0.9726 | 0.0000 | 0.2874 |
| | | SAL | 0.7756 | 0.7820 | 0.7692 | 0.0064 | 0.5733 | 0.9764 | 0.0000 | 0.2804 |
| | | STL | 0.6482 | 0.7712 | 0.5251 | 0.1230 | 0.5736 | 0.9719 | 0.0000 | 0.2658 |
| | 4 | RAN | 0.7715 | 0.7715 | 0.7715 | 0.0000 | 0.5690 | 0.9714 | 0.0000 | 0.2787 |
| | | SAL | 0.7756 | 0.7808 | 0.7704 | 0.0052 | 0.5748 | 0.9742 | 0.0000 | 0.2780 |
| | | STL | 0.6480 | 0.7723 | 0.5237 | 0.1243 | 0.5763 | 0.9713 | 0.0000 | 0.2661 |

Table E.23: grouped-liblinear-enron-GB3-ALL-ALL-75

| GB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | Group Type | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | RAN | 0.7472 | 0.7472 | 0.7472 | 0.0000 | 0.5470 | 0.9701 | 0.0000 | 0.2839 |
| | | SAL | 0.7440 | 0.7440 | 0.7440 | 0.0000 | 0.5326 | 0.9711 | 0.0000 | 0.2882 |
| | | STL | 0.7440 | 0.7440 | 0.7440 | 0.0000 | 0.5326 | 0.9711 | 0.0000 | 0.2882 |
| | 1 | RAN | 0.7453 | 0.7453 | 0.7453 | 0.0000 | 0.5413 | 0.9700 | 0.0000 | 0.2850 |
| | | SAL | 0.7426 | 0.7426 | 0.7426 | 0.0000 | 0.5358 | 0.9867 | 0.0000 | 0.2903 |
| | | STL | 0.7426 | 0.7426 | 0.7426 | 0.0000 | 0.5358 | 0.9867 | 0.0000 | 0.2903 |
| | 2 | RAN | 0.7432 | 0.7432 | 0.7432 | 0.0000 | 0.5451 | 0.9666 | 0.0000 | 0.2813 |
| | | SAL | 0.7417 | 0.7417 | 0.7417 | 0.0000 | 0.5376 | 0.9867 | 0.0000 | 0.2848 |
| | | STL | 0.7417 | 0.7417 | 0.7417 | 0.0000 | 0.5376 | 0.9867 | 0.0000 | 0.2848 |

Table E.24: grouped-liblinear-enron-GB3-ALL-ALL-150

| OSB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | RAN | 0.8946 | 0.9696 | 0.6599 | 0.0718 | 0.7269 | 1.0000 | 0.0000 | 0.2605 |
| | | SAL | 0.9147 | 0.9732 | 0.7892 | 0.0337 | 0.7246 | 0.9903 | 0.0000 | 0.2633 |
| | | STL | 0.7979 | 0.9579 | 0.5185 | 0.1085 | 0.7643 | 0.9841 | 0.0000 | 0.1994 |
| | 1 | RAN | 0.8970 | 0.9619 | 0.6565 | 0.0575 | 0.7324 | 0.9869 | 0.0000 | 0.2481 |
| | | SAL | 0.9147 | 0.9673 | 0.7777 | 0.0410 | 0.7227 | 0.9921 | 0.0000 | 0.2643 |
| | | STL | 0.7885 | 0.9741 | 0.5185 | 0.1165 | 0.7573 | 0.9853 | 0.0000 | 0.2004 |
| | 2 | RAN | 0.8904 | 0.9765 | 0.6446 | 0.0710 | 0.7325 | 0.9923 | 0.0000 | 0.2548 |
| | | SAL | 0.9142 | 0.9673 | 0.7812 | 0.0387 | 0.7203 | 0.9877 | 0.0000 | 0.2642 |
| | | STL | 0.7891 | 0.9740 | 0.5185 | 0.1187 | 0.7579 | 0.9851 | 0.0000 | 0.2020 |
| | 4 | RAN | 0.9027 | 0.9762 | 0.7346 | 0.0487 | 0.7313 | 0.9884 | 0.0000 | 0.2534 |
| | | SAL | 0.9142 | 0.9677 | 0.7777 | 0.0392 | 0.7208 | 0.9877 | 0.0000 | 0.2634 |
| | | STL | 0.7893 | 0.9738 | 0.5185 | 0.1187 | 0.7582 | 0.9856 | 0.0000 | 0.2019 |
| | 8 | RAN | 0.8876 | 0.9706 | 0.7021 | 0.0675 | 0.7249 | 0.9840 | 0.0000 | 0.2584 |
| | | SAL | 0.9144 | 0.9673 | 0.7777 | 0.0396 | 0.7207 | 0.9877 | 0.0000 | 0.2642 |
| | | STL | 0.7889 | 0.9740 | 0.5185 | 0.1191 | 0.7581 | 0.9853 | 0.0000 | 0.2020 |
| | 16 | RAN | 0.8916 | 0.9848 | 0.7284 | 0.0649 | 0.7327 | 0.9924 | 0.0000 | 0.2513 |
| | | SAL | 0.9143 | 0.9677 | 0.7777 | 0.0392 | 0.7206 | 0.9877 | 0.0000 | 0.2640 |
| | | STL | 0.7889 | 0.9744 | 0.5185 | 0.1185 | 0.7578 | 0.9853 | 0.0000 | 0.2017 |

Table E.25: grouped-liblinear-enron-OSB3-ALL-ALL-5

| OSB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | Group Type | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | RAN | 0.8643 | 0.9469 | 0.7573 | 0.0505 | 0.6814 | 0.9867 | 0.0000 | 0.2595 |
| | | SAL | 0.8758 | 0.9462 | 0.7660 | 0.0419 | 0.6788 | 0.9867 | 0.0000 | 0.2647 |
| | | STL | 0.7349 | 0.9353 | 0.5106 | 0.1231 | 0.7054 | 0.9841 | 0.0000 | 0.2236 |
| | 1 | RAN | 0.8669 | 0.9516 | 0.7406 | 0.0518 | 0.6841 | 0.9849 | 0.0000 | 0.2616 |
| | | SAL | 0.8813 | 0.9316 | 0.7407 | 0.0470 | 0.6847 | 0.9888 | 0.0000 | 0.2642 |
| | | STL | 0.7405 | 0.9363 | 0.5319 | 0.1154 | 0.7122 | 0.9844 | 0.0000 | 0.2126 |
| | 2 | RAN | 0.8638 | 0.9432 | 0.7652 | 0.0573 | 0.6847 | 1.0000 | 0.0000 | 0.2616 |
| | | SAL | 0.8801 | 0.9319 | 0.7410 | 0.0459 | 0.6804 | 0.9870 | 0.0000 | 0.2642 |
| | | STL | 0.7401 | 0.9368 | 0.5319 | 0.1174 | 0.7120 | 0.9817 | 0.0000 | 0.2119 |
| | 4 | RAN | 0.8600 | 0.9446 | 0.7783 | 0.0549 | 0.6809 | 0.9854 | 0.0000 | 0.2616 |
| | | SAL | 0.8804 | 0.9319 | 0.7448 | 0.0450 | 0.6814 | 0.9870 | 0.0000 | 0.2628 |
| | | STL | 0.7393 | 0.9369 | 0.5319 | 0.1172 | 0.7114 | 0.9812 | 0.0000 | 0.2121 |
| | 8 | RAN | 0.8652 | 0.9482 | 0.7428 | 0.0554 | 0.6859 | 0.9851 | 0.0000 | 0.2573 |
| | | SAL | 0.8793 | 0.9317 | 0.7381 | 0.0450 | 0.6782 | 0.9870 | 0.0000 | 0.2670 |
| | | STL | 0.7401 | 0.9372 | 0.5319 | 0.1174 | 0.7119 | 0.9817 | 0.0000 | 0.2121 |

Table E.26: grouped-liblinear-enron-OSB3-ALL-ALL-10

| OSB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | Group Type | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | RAN | 0.8300 | 0.8954 | 0.7736 | 0.0457 | 0.6303 | 0.9815 | 0.0000 | 0.2729 |
| | | SAL | 0.8320 | 0.8928 | 0.7447 | 0.0480 | 0.6334 | 0.9810 | 0.0000 | 0.2650 |
| | | STL | 0.6858 | 0.8664 | 0.5385 | 0.1179 | 0.6486 | 0.9771 | 0.0000 | 0.2403 |
| | 1 | RAN | 0.8262 | 0.8850 | 0.7908 | 0.0382 | 0.6321 | 0.9867 | 0.0000 | 0.2703 |
| | | SAL | 0.8379 | 0.8634 | 0.8210 | 0.0132 | 0.6345 | 0.9801 | 0.0000 | 0.2718 |
| | | STL | 0.6823 | 0.8636 | 0.5350 | 0.1156 | 0.6498 | 0.9752 | 0.0000 | 0.2362 |
| | 2 | RAN | 0.8349 | 0.8516 | 0.8018 | 0.0171 | 0.6336 | 1.0000 | 0.0000 | 0.2708 |
| | | SAL | 0.8388 | 0.8775 | 0.8140 | 0.0207 | 0.6377 | 0.9787 | 0.0000 | 0.2689 |
| | | STL | 0.6806 | 0.8630 | 0.5350 | 0.1162 | 0.6485 | 0.9769 | 0.0000 | 0.2372 |

Table E.27: grouped-liblinear-enron-OSB3-ALL-ALL-25

| OSB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | Group Type | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | RAN | 0.8012 | 0.8280 | 0.7684 | 0.0247 | 0.5999 | 0.9727 | 0.0000 | 0.2719 |
| | | SAL | 0.8063 | 0.8569 | 0.7447 | 0.0465 | 0.6051 | 0.9793 | 0.0000 | 0.2726 |
| | | STL | 0.6625 | 0.8147 | 0.5160 | 0.1220 | 0.6119 | 0.9785 | 0.0000 | 0.2481 |
| | 1 | RAN | 0.8024 | 0.8343 | 0.7672 | 0.0275 | 0.5976 | 0.9717 | 0.0000 | 0.2769 |
| | | SAL | 0.8025 | 0.8219 | 0.7801 | 0.0172 | 0.6010 | 0.9764 | 0.0000 | 0.2706 |
| | | STL | 0.6575 | 0.8132 | 0.5126 | 0.1230 | 0.6096 | 0.9771 | 0.0000 | 0.2493 |

Table E.28: grouped-liblinear-enron-OSB3-ALL-ALL-50

| OSB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | Group Type | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | RAN | 0.7842 | 0.7888 | 0.7796 | 0.0046 | 0.5802 | 0.9741 | 0.0000 | 0.2784 |
| | | SAL | 0.7852 | 0.7931 | 0.7774 | 0.0079 | 0.5874 | 0.9786 | 0.0000 | 0.2708 |
| | | STL | 0.6715 | 0.7884 | 0.5547 | 0.1169 | 0.5898 | 0.9766 | 0.0000 | 0.2615 |
| | 1 | RAN | 0.7832 | 0.8011 | 0.7653 | 0.0179 | 0.5876 | 0.9705 | 0.0000 | 0.2710 |
| | | SAL | 0.7859 | 0.7894 | 0.7823 | 0.0035 | 0.5870 | 0.9751 | 0.0000 | 0.2726 |
| | | STL | 0.6678 | 0.7898 | 0.5458 | 0.1220 | 0.5950 | 0.9758 | 0.0000 | 0.2582 |

Table E.29: grouped-liblinear-enron-OSB3-ALL-ALL-75

| OSB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | Group Type | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | RAN | 0.7553 | 0.7553 | 0.7553 | 0.0000 | 0.5537 | 0.9754 | 0.0000 | 0.2782 |
| | | SAL | 0.7543 | 0.7543 | 0.7543 | 0.0000 | 0.5563 | 0.9716 | 0.0000 | 0.2778 |
| | | STL | 0.7543 | 0.7543 | 0.7543 | 0.0000 | 0.5563 | 0.9716 | 0.0000 | 0.2778 |

Table E.30: grouped-liblinear-enron-OSB3-ALL-ALL-150

# APPENDIX F:
# Grouped LibLinear Results for the Twitter Short Message Corpus

| GM1 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | Group Type | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | RAN | 0.6283 | 0.8081 | 0.4918 | 0.0706 | 0.6077 | 0.9696 | 0.1892 | 0.1404 |
| | | SAL | 0.6321 | 0.7631 | 0.5068 | 0.0720 | 0.6100 | 0.9363 | 0.2195 | 0.1374 |
| | | STL | 0.6032 | 0.8089 | 0.4737 | 0.0758 | 0.5893 | 0.9391 | 0.1791 | 0.1458 |
| | 1 | RAN | 0.6333 | 0.8037 | 0.5031 | 0.0723 | 0.6074 | 0.9588 | 0.1449 | 0.1479 |
| | | SAL | 0.6259 | 0.8211 | 0.5083 | 0.0802 | 0.6013 | 0.9597 | 0.1429 | 0.1523 |
| | | STL | 0.6093 | 0.8117 | 0.4713 | 0.0787 | 0.6015 | 0.9347 | 0.1690 | 0.1326 |
| | 2 | RAN | 0.6223 | 0.8966 | 0.4754 | 0.0969 | 0.6070 | 0.9602 | 0.2000 | 0.1431 |
| | | SAL | 0.6025 | 0.8093 | 0.4441 | 0.0877 | 0.5720 | 0.9697 | 0.1509 | 0.1636 |
| | | STL | 0.6195 | 0.8049 | 0.4218 | 0.0796 | 0.6055 | 0.9328 | 0.0000 | 0.1442 |
| | 4 | RAN | 0.6126 | 0.7957 | 0.3846 | 0.0918 | 0.5914 | 0.9811 | 0.1200 | 0.1514 |
| | | SAL | 0.6266 | 0.8546 | 0.4823 | 0.0837 | 0.6041 | 0.9560 | 0.1695 | 0.1486 |
| | | STL | 0.6123 | 0.8062 | 0.4659 | 0.0781 | 0.6021 | 0.9347 | 0.1972 | 0.1342 |
| | 8 | RAN | 0.6356 | 0.8084 | 0.4848 | 0.0714 | 0.6088 | 0.9718 | 0.1613 | 0.1491 |
| | | SAL | 0.6321 | 0.9026 | 0.4828 | 0.0909 | 0.6121 | 0.9735 | 0.1404 | 0.1448 |
| | | STL | 0.6143 | 0.8117 | 0.4661 | 0.0789 | 0.6052 | 0.9347 | 0.1972 | 0.1310 |
| | 16 | RAN | 0.6179 | 0.8112 | 0.4645 | 0.0902 | 0.5989 | 0.9470 | 0.1818 | 0.1455 |
| | | SAL | 0.6219 | 0.8324 | 0.4563 | 0.0781 | 0.6022 | 0.9600 | 0.2182 | 0.1413 |
| | | STL | 0.6145 | 0.8062 | 0.4458 | 0.0756 | 0.6028 | 0.9347 | 0.1515 | 0.1283 |

Table F.1: grouped-liblinear-twitter-GM1-ALL-ALL-5

| GM1 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | RAN | 0.4768 | 0.6234 | 0.3448 | 0.0817 | 0.4500 | 0.9556 | 0.0702 | 0.1651 |
| | | SAL | 0.4978 | 0.5788 | 0.3952 | 0.0572 | 0.4670 | 0.9527 | 0.0385 | 0.1615 |
| | | STL | 0.4540 | 0.6063 | 0.3578 | 0.0697 | 0.4443 | 0.9389 | 0.1250 | 0.1545 |
| | 1 | RAN | 0.4943 | 0.5836 | 0.3936 | 0.0629 | 0.4682 | 0.9482 | 0.0755 | 0.1626 |
| | | SAL | 0.4773 | 0.6389 | 0.3627 | 0.0640 | 0.4545 | 0.9405 | 0.0923 | 0.1607 |
| | | STL | 0.4724 | 0.6329 | 0.3766 | 0.0695 | 0.4587 | 0.9384 | 0.0889 | 0.1539 |
| | 2 | RAN | 0.4914 | 0.5881 | 0.3884 | 0.0627 | 0.4631 | 0.9513 | 0.0833 | 0.1706 |
| | | SAL | 0.4859 | 0.6567 | 0.3888 | 0.0779 | 0.4582 | 0.9699 | 0.0702 | 0.1608 |
| | | STL | 0.4761 | 0.6321 | 0.3358 | 0.0673 | 0.4638 | 0.9381 | 0.1446 | 0.1524 |
| | 4 | RAN | 0.4947 | 0.6310 | 0.3849 | 0.0544 | 0.4729 | 0.9517 | 0.1250 | 0.1570 |
| | | SAL | 0.4853 | 0.6900 | 0.3778 | 0.0916 | 0.4542 | 0.9487 | 0.0714 | 0.1781 |
| | | STL | 0.4722 | 0.6362 | 0.3184 | 0.0728 | 0.4613 | 0.9402 | 0.0299 | 0.1551 |
| | 8 | RAN | 0.4946 | 0.6282 | 0.3895 | 0.0632 | 0.4639 | 0.9472 | 0.0000 | 0.1622 |
| | | SAL | 0.4787 | 0.7194 | 0.3994 | 0.0862 | 0.4495 | 0.9621 | 0.1034 | 0.1636 |
| | | STL | 0.4714 | 0.6362 | 0.3080 | 0.0821 | 0.4593 | 0.9402 | 0.0400 | 0.1648 |
| | 16 | RAN | 0.4904 | 0.6124 | 0.3844 | 0.0675 | 0.4600 | 0.9363 | 0.0769 | 0.1604 |
| | | SAL | 0.4817 | 0.5909 | 0.3985 | 0.0591 | 0.4534 | 0.9474 | 0.0370 | 0.1659 |
| | | STL | 0.4679 | 0.6362 | 0.2846 | 0.0859 | 0.4568 | 0.9402 | 0.1026 | 0.1576 |

Table F.2: grouped-liblinear-twitter-GM1-ALL-ALL-10

| GM1 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | RAN | 0.3441 | 0.4816 | 0.2912 | 0.0644 | 0.3144 | 0.9344 | 0.0000 | 0.1658 |
| | | SAL | 0.3544 | 0.4245 | 0.2735 | 0.0484 | 0.3248 | 0.8820 | 0.0000 | 0.1725 |
| | | STL | 0.3399 | 0.4550 | 0.2916 | 0.0538 | 0.3199 | 0.8806 | 0.0000 | 0.1707 |
| | 1 | RAN | 0.3377 | 0.4119 | 0.2408 | 0.0661 | 0.3121 | 0.9167 | 0.0000 | 0.1736 |
| | | SAL | 0.3531 | 0.4352 | 0.3091 | 0.0518 | 0.3188 | 0.9221 | 0.0000 | 0.1782 |
| | | STL | 0.3318 | 0.4390 | 0.2660 | 0.0582 | 0.3170 | 0.9052 | 0.0000 | 0.1699 |
| | 2 | RAN | 0.3535 | 0.4084 | 0.3176 | 0.0279 | 0.3207 | 0.8960 | 0.0377 | 0.1672 |
| | | SAL | 0.3477 | 0.4309 | 0.2878 | 0.0500 | 0.3197 | 0.9225 | 0.0000 | 0.1779 |
| | | STL | 0.3382 | 0.4245 | 0.2714 | 0.0463 | 0.3271 | 0.8969 | 0.0000 | 0.1696 |
| | 4 | RAN | 0.3566 | 0.4015 | 0.3092 | 0.0320 | 0.3249 | 0.9119 | 0.0000 | 0.1747 |
| | | SAL | 0.3486 | 0.4402 | 0.2859 | 0.0583 | 0.3213 | 0.9358 | 0.0000 | 0.1729 |
| | | STL | 0.3478 | 0.4372 | 0.2811 | 0.0482 | 0.3331 | 0.8851 | 0.0000 | 0.1694 |
| | 8 | RAN | 0.3442 | 0.4118 | 0.3082 | 0.0327 | 0.3166 | 0.9051 | 0.0000 | 0.1671 |
| | | SAL | 0.3520 | 0.4186 | 0.2944 | 0.0498 | 0.3267 | 0.9231 | 0.0000 | 0.1788 |
| | | STL | 0.3294 | 0.4296 | 0.2591 | 0.0566 | 0.3135 | 0.8851 | 0.0000 | 0.1656 |
| | 16 | RAN | 0.3482 | 0.3976 | 0.3137 | 0.0335 | 0.3136 | 0.9011 | 0.0000 | 0.1701 |
| | | SAL | 0.3375 | 0.4094 | 0.2879 | 0.0373 | 0.3074 | 0.8981 | 0.0000 | 0.1742 |
| | | STL | 0.3376 | 0.4296 | 0.2651 | 0.0514 | 0.3188 | 0.8851 | 0.0000 | 0.1670 |

Table F.3: grouped-liblinear-twitter-GM1-ALL-ALL-25

| GM1 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | Group Type | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | RAN | 0.2776 | 0.3371 | 0.2451 | 0.0421 | 0.2566 | 0.8780 | 0.0000 | 0.1768 |
| | | SAL | 0.2659 | 0.2759 | 0.2510 | 0.0107 | 0.2405 | 0.8759 | 0.0000 | 0.1688 |
| | | STL | 0.2645 | 0.3273 | 0.2190 | 0.0459 | 0.2518 | 0.8798 | 0.0000 | 0.1799 |
| | 1 | RAN | 0.2736 | 0.3120 | 0.2219 | 0.0380 | 0.2521 | 0.9153 | 0.0000 | 0.1736 |
| | | SAL | 0.2769 | 0.3173 | 0.2438 | 0.0305 | 0.2506 | 0.8788 | 0.0000 | 0.1745 |
| | | STL | 0.2606 | 0.3142 | 0.2249 | 0.0386 | 0.2421 | 0.8647 | 0.0000 | 0.1758 |
| | 2 | RAN | 0.2741 | 0.3248 | 0.2180 | 0.0438 | 0.2398 | 0.8905 | 0.0000 | 0.1723 |
| | | SAL | 0.2644 | 0.3100 | 0.2250 | 0.0350 | 0.2351 | 0.8922 | 0.0000 | 0.1718 |
| | | STL | 0.2731 | 0.3338 | 0.2284 | 0.0445 | 0.2510 | 0.8838 | 0.0000 | 0.1814 |
| | 4 | RAN | 0.2826 | 0.3236 | 0.2287 | 0.0398 | 0.2524 | 0.8766 | 0.0000 | 0.1675 |
| | | SAL | 0.2717 | 0.3272 | 0.2032 | 0.0515 | 0.2452 | 0.8889 | 0.0000 | 0.1676 |
| | | STL | 0.2586 | 0.3155 | 0.2162 | 0.0418 | 0.2416 | 0.8880 | 0.0000 | 0.1793 |
| | 8 | RAN | 0.2773 | 0.2996 | 0.2637 | 0.0159 | 0.2504 | 0.9119 | 0.0000 | 0.1770 |
| | | SAL | 0.2742 | 0.2924 | 0.2446 | 0.0211 | 0.2381 | 0.8449 | 0.0000 | 0.1686 |
| | | STL | 0.2623 | 0.3171 | 0.2274 | 0.0393 | 0.2409 | 0.9010 | 0.0000 | 0.1722 |
| | 16 | RAN | 0.2759 | 0.2963 | 0.2501 | 0.0192 | 0.2525 | 0.9035 | 0.0000 | 0.1689 |
| | | SAL | 0.2698 | 0.2920 | 0.2504 | 0.0171 | 0.2370 | 0.8832 | 0.0000 | 0.1713 |
| | | STL | 0.2672 | 0.3326 | 0.2344 | 0.0463 | 0.2480 | 0.8906 | 0.0000 | 0.1733 |

Table F.4: grouped-liblinear-twitter-GM1-ALL-ALL-50

| GM1 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | RAN | 0.2396 | 0.2738 | 0.2054 | 0.0342 | 0.2200 | 0.8390 | 0.0000 | 0.1619 |
| | | SAL | 0.2165 | 0.2521 | 0.1810 | 0.0356 | 0.1989 | 0.8321 | 0.0000 | 0.1745 |
| | | STL | 0.2257 | 0.2633 | 0.1881 | 0.0376 | 0.2112 | 0.8687 | 0.0000 | 0.1765 |
| | 1 | RAN | 0.2337 | 0.2554 | 0.2120 | 0.0217 | 0.2063 | 0.8630 | 0.0000 | 0.1679 |
| | | SAL | 0.2347 | 0.2595 | 0.2098 | 0.0249 | 0.2140 | 0.9157 | 0.0000 | 0.1727 |
| | | STL | 0.2197 | 0.2626 | 0.1768 | 0.0429 | 0.2074 | 0.8435 | 0.0000 | 0.1746 |
| | 2 | RAN | 0.2313 | 0.2641 | 0.1986 | 0.0327 | 0.2093 | 0.8722 | 0.0000 | 0.1780 |
| | | SAL | 0.2367 | 0.2648 | 0.2086 | 0.0281 | 0.2173 | 0.8417 | 0.0000 | 0.1683 |
| | | STL | 0.2274 | 0.2730 | 0.1818 | 0.0456 | 0.2114 | 0.8750 | 0.0000 | 0.1723 |
| | 4 | RAN | 0.2299 | 0.2328 | 0.2271 | 0.0028 | 0.2063 | 0.8971 | 0.0000 | 0.1665 |
| | | SAL | 0.2391 | 0.2776 | 0.2006 | 0.0385 | 0.2150 | 0.8387 | 0.0000 | 0.1719 |
| | | STL | 0.2240 | 0.2605 | 0.1874 | 0.0365 | 0.2091 | 0.8343 | 0.0000 | 0.1674 |
| | 8 | RAN | 0.2451 | 0.2487 | 0.2416 | 0.0035 | 0.2157 | 0.8873 | 0.0000 | 0.1741 |
| | | SAL | 0.2413 | 0.2910 | 0.1915 | 0.0498 | 0.2100 | 0.8429 | 0.0000 | 0.1686 |
| | | STL | 0.2199 | 0.2654 | 0.1743 | 0.0456 | 0.2033 | 0.8116 | 0.0000 | 0.1729 |
| | 16 | RAN | 0.2372 | 0.2745 | 0.2000 | 0.0372 | 0.2064 | 0.9037 | 0.0000 | 0.1756 |
| | | SAL | 0.2387 | 0.2639 | 0.2136 | 0.0252 | 0.2094 | 0.9034 | 0.0000 | 0.1691 |
| | | STL | 0.2345 | 0.2780 | 0.1910 | 0.0435 | 0.2114 | 0.8621 | 0.0000 | 0.1771 |

Table F.5: grouped-liblinear-twitter-GM1-ALL-ALL-75

| GM1 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | RAN | 0.1802 | 0.1802 | 0.1802 | 0.0000 | 0.1609 | 0.8582 | 0.0000 | 0.1599 |
| | | SAL | 0.1875 | 0.1875 | 0.1875 | 0.0000 | 0.1640 | 0.8127 | 0.0000 | 0.1634 |
| | | STL | 0.1875 | 0.1875 | 0.1875 | 0.0000 | 0.1640 | 0.8127 | 0.0000 | 0.1634 |
| | 1 | RAN | 0.1802 | 0.1802 | 0.1802 | 0.0000 | 0.1554 | 0.8071 | 0.0000 | 0.1605 |
| | | SAL | 0.1932 | 0.1932 | 0.1932 | 0.0000 | 0.1687 | 0.8212 | 0.0000 | 0.1650 |
| | | STL | 0.1932 | 0.1932 | 0.1932 | 0.0000 | 0.1687 | 0.8212 | 0.0000 | 0.1650 |
| | 2 | RAN | 0.1821 | 0.1821 | 0.1821 | 0.0000 | 0.1617 | 0.7698 | 0.0000 | 0.1628 |
| | | SAL | 0.1833 | 0.1833 | 0.1833 | 0.0000 | 0.1585 | 0.7807 | 0.0000 | 0.1548 |
| | | STL | 0.1833 | 0.1833 | 0.1833 | 0.0000 | 0.1585 | 0.7807 | 0.0000 | 0.1548 |
| | 4 | RAN | 0.1943 | 0.1943 | 0.1943 | 0.0000 | 0.1702 | 0.8792 | 0.0000 | 0.1668 |
| | | SAL | 0.1910 | 0.1910 | 0.1910 | 0.0000 | 0.1647 | 0.7697 | 0.0000 | 0.1556 |
| | | STL | 0.1910 | 0.1910 | 0.1910 | 0.0000 | 0.1647 | 0.7697 | 0.0000 | 0.1556 |
| | 8 | RAN | 0.1913 | 0.1913 | 0.1913 | 0.0000 | 0.1649 | 0.8139 | 0.0000 | 0.1687 |
| | | SAL | 0.1884 | 0.1884 | 0.1884 | 0.0000 | 0.1636 | 0.7644 | 0.0000 | 0.1631 |
| | | STL | 0.1884 | 0.1884 | 0.1884 | 0.0000 | 0.1636 | 0.7644 | 0.0000 | 0.1631 |
| | 16 | RAN | 0.1865 | 0.1865 | 0.1865 | 0.0000 | 0.1592 | 0.8143 | 0.0000 | 0.1666 |
| | | SAL | 0.1883 | 0.1883 | 0.1883 | 0.0000 | 0.1671 | 0.7732 | 0.0000 | 0.1588 |
| | | STL | 0.1883 | 0.1883 | 0.1883 | 0.0000 | 0.1671 | 0.7732 | 0.0000 | 0.1588 |

Table F.6: grouped-liblinear-twitter-GM1-ALL-ALL-150

| GM2 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | RAN | 0.4955 | 0.7162 | 0.3041 | 0.0815 | 0.4553 | 0.9272 | 0.0000 | 0.1764 |
| | | SAL | 0.4834 | 0.7282 | 0.2664 | 0.1007 | 0.4476 | 0.9272 | 0.0000 | 0.1688 |
| | | STL | 0.4744 | 0.6572 | 0.3283 | 0.0813 | 0.4474 | 0.8897 | 0.0000 | 0.1570 |
| | 1 | RAN | 0.6339 | 0.8454 | 0.4772 | 0.0813 | 0.6069 | 0.9886 | 0.1159 | 0.1596 |
| | | SAL | 0.6252 | 0.7790 | 0.4664 | 0.0795 | 0.6051 | 0.9771 | 0.1739 | 0.1438 |
| | | STL | 0.6133 | 0.7656 | 0.4881 | 0.0714 | 0.5967 | 0.9202 | 0.1250 | 0.1404 |
| | 2 | RAN | 0.6219 | 0.8172 | 0.4335 | 0.0934 | 0.6025 | 0.9697 | 0.1071 | 0.1478 |
| | | SAL | 0.6324 | 0.7629 | 0.3934 | 0.0813 | 0.6130 | 0.9524 | 0.2143 | 0.1385 |
| | | STL | 0.6120 | 0.8232 | 0.4517 | 0.0838 | 0.6006 | 0.9202 | 0.2716 | 0.1400 |
| | 4 | RAN | 0.6420 | 0.8157 | 0.5014 | 0.0783 | 0.6202 | 0.9545 | 0.2258 | 0.1400 |
| | | SAL | 0.6196 | 0.8210 | 0.4820 | 0.0810 | 0.5996 | 0.9488 | 0.1951 | 0.1453 |
| | | STL | 0.6144 | 0.7669 | 0.4245 | 0.0751 | 0.6035 | 0.9202 | 0.1639 | 0.1351 |
| | 8 | RAN | 0.6415 | 0.7996 | 0.5168 | 0.0782 | 0.6205 | 0.9732 | 0.2326 | 0.1453 |
| | | SAL | 0.6309 | 0.8489 | 0.4951 | 0.0678 | 0.6136 | 0.9773 | 0.1818 | 0.1390 |
| | | STL | 0.6123 | 0.7669 | 0.4773 | 0.0707 | 0.6048 | 0.9202 | 0.2667 | 0.1265 |
| | 16 | RAN | 0.6358 | 0.8544 | 0.4986 | 0.0831 | 0.6204 | 0.9603 | 0.2985 | 0.1281 |
| | | SAL | 0.6272 | 0.8015 | 0.4094 | 0.1010 | 0.6102 | 0.9579 | 0.2157 | 0.1468 |
| | | STL | 0.6144 | 0.7870 | 0.5018 | 0.0758 | 0.6051 | 0.9202 | 0.2368 | 0.1282 |

Table F.7: grouped-liblinear-twitter-GM2-ALL-ALL-5

| GM2 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | Group Type | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | RAN | 0.3685 | 0.5557 | 0.1985 | 0.0875 | 0.3390 | 0.9150 | 0.0000 | 0.1702 |
| | | SAL | 0.3721 | 0.5725 | 0.2590 | 0.0864 | 0.3469 | 0.8176 | 0.0000 | 0.1686 |
| | | STL | 0.3695 | 0.4854 | 0.2646 | 0.0539 | 0.3506 | 0.8539 | 0.0000 | 0.1669 |
| | 1 | RAN | 0.5044 | 0.6219 | 0.3851 | 0.0668 | 0.4737 | 0.9575 | 0.0357 | 0.1671 |
| | | SAL | 0.4832 | 0.5897 | 0.3429 | 0.0674 | 0.4514 | 0.9531 | 0.0000 | 0.1722 |
| | | STL | 0.4836 | 0.6154 | 0.3430 | 0.0631 | 0.4721 | 0.8945 | 0.1667 | 0.1498 |
| | 2 | RAN | 0.4934 | 0.5992 | 0.3683 | 0.0549 | 0.4690 | 0.9509 | 0.0923 | 0.1542 |
| | | SAL | 0.5073 | 0.6491 | 0.4107 | 0.0599 | 0.4802 | 0.9549 | 0.0923 | 0.1667 |
| | | STL | 0.4700 | 0.6213 | 0.3560 | 0.0672 | 0.4575 | 0.8963 | 0.0879 | 0.1536 |
| | 4 | RAN | 0.4989 | 0.6303 | 0.3851 | 0.0732 | 0.4696 | 0.9771 | 0.0303 | 0.1692 |
| | | SAL | 0.5058 | 0.6924 | 0.4116 | 0.0731 | 0.4763 | 0.9363 | 0.0000 | 0.1648 |
| | | STL | 0.4840 | 0.6013 | 0.3379 | 0.0650 | 0.4676 | 0.8928 | 0.0714 | 0.1573 |
| | 8 | RAN | 0.4895 | 0.6693 | 0.3939 | 0.0782 | 0.4646 | 0.9524 | 0.1463 | 0.1658 |
| | | SAL | 0.4870 | 0.5608 | 0.4050 | 0.0400 | 0.4583 | 0.9502 | 0.1042 | 0.1627 |
| | | STL | 0.4762 | 0.6163 | 0.3593 | 0.0619 | 0.4638 | 0.8980 | 0.0857 | 0.1519 |
| | 16 | RAN | 0.4965 | 0.6961 | 0.3142 | 0.0889 | 0.4709 | 0.9421 | 0.0000 | 0.1724 |
| | | SAL | 0.4962 | 0.6383 | 0.3804 | 0.0677 | 0.4714 | 0.9358 | 0.1111 | 0.1564 |
| | | STL | 0.4746 | 0.6163 | 0.3733 | 0.0591 | 0.4628 | 0.8980 | 0.1379 | 0.1440 |

Table F.8: grouped-liblinear-twitter-GM2-ALL-ALL-10

| GM2 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | RAN | 0.2618 | 0.3285 | 0.2022 | 0.0403 | 0.2437 | 0.8399 | 0.0000 | 0.1704 |
| | | SAL | 0.2694 | 0.3163 | 0.2364 | 0.0237 | 0.2376 | 0.7742 | 0.0000 | 0.1790 |
| | | STL | 0.2613 | 0.3004 | 0.1882 | 0.0369 | 0.2498 | 0.8473 | 0.0000 | 0.1706 |
| | 1 | RAN | 0.3556 | 0.3991 | 0.3017 | 0.0342 | 0.3266 | 0.9299 | 0.0000 | 0.1729 |
| | | SAL | 0.3512 | 0.4945 | 0.2926 | 0.0698 | 0.3227 | 0.9037 | 0.0000 | 0.1790 |
| | | STL | 0.3338 | 0.4079 | 0.2702 | 0.0418 | 0.3186 | 0.8871 | 0.0000 | 0.1689 |
| | 2 | RAN | 0.3525 | 0.4076 | 0.2881 | 0.0379 | 0.3219 | 0.8880 | 0.0000 | 0.1689 |
| | | SAL | 0.3555 | 0.4433 | 0.2642 | 0.0587 | 0.3285 | 0.9542 | 0.0000 | 0.1683 |
| | | STL | 0.3313 | 0.4132 | 0.2655 | 0.0500 | 0.3154 | 0.8618 | 0.0241 | 0.1658 |
| | 4 | RAN | 0.3577 | 0.4149 | 0.3246 | 0.0338 | 0.3291 | 0.8548 | 0.0000 | 0.1675 |
| | | SAL | 0.3598 | 0.4450 | 0.3107 | 0.0441 | 0.3372 | 0.9438 | 0.0345 | 0.1718 |
| | | STL | 0.3403 | 0.4418 | 0.2665 | 0.0576 | 0.3249 | 0.8463 | 0.0000 | 0.1681 |
| | 8 | RAN | 0.3551 | 0.4446 | 0.2818 | 0.0548 | 0.3224 | 0.9011 | 0.0000 | 0.1673 |
| | | SAL | 0.3533 | 0.4469 | 0.3027 | 0.0480 | 0.3249 | 0.9136 | 0.0000 | 0.1711 |
| | | STL | 0.3253 | 0.4172 | 0.2535 | 0.0541 | 0.3132 | 0.8767 | 0.0682 | 0.1608 |
| | 16 | RAN | 0.3463 | 0.4218 | 0.2916 | 0.0476 | 0.3175 | 0.9213 | 0.0000 | 0.1831 |
| | | SAL | 0.3572 | 0.4682 | 0.2755 | 0.0805 | 0.3310 | 0.8782 | 0.0313 | 0.1664 |
| | | STL | 0.3276 | 0.4381 | 0.2419 | 0.0648 | 0.3166 | 0.8436 | 0.0000 | 0.1707 |

Table F.9: grouped-liblinear-twitter-GM2-ALL-ALL-25

| GM2 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | RAN | 0.2089 | 0.2220 | 0.1947 | 0.0112 | 0.1974 | 0.8315 | 0.0000 | 0.1739 |
| | | SAL | 0.2107 | 0.2265 | 0.1860 | 0.0177 | 0.2008 | 0.8333 | 0.0000 | 0.1720 |
| | | STL | 0.2095 | 0.2745 | 0.1587 | 0.0484 | 0.2034 | 0.8333 | 0.0000 | 0.1736 |
| | 1 | RAN | 0.2702 | 0.3381 | 0.2310 | 0.0482 | 0.2478 | 0.8592 | 0.0000 | 0.1705 |
| | | SAL | 0.2781 | 0.3185 | 0.2483 | 0.0296 | 0.2458 | 0.8750 | 0.0000 | 0.1685 |
| | | STL | 0.2555 | 0.3090 | 0.2275 | 0.0378 | 0.2424 | 0.8661 | 0.0000 | 0.1744 |
| | 2 | RAN | 0.2701 | 0.2863 | 0.2559 | 0.0125 | 0.2423 | 0.8521 | 0.0000 | 0.1803 |
| | | SAL | 0.2794 | 0.3221 | 0.2116 | 0.0484 | 0.2533 | 0.8777 | 0.0000 | 0.1780 |
| | | STL | 0.2625 | 0.3178 | 0.2331 | 0.0391 | 0.2454 | 0.8571 | 0.0000 | 0.1802 |
| | 4 | RAN | 0.2742 | 0.3019 | 0.2591 | 0.0196 | 0.2442 | 0.8473 | 0.0000 | 0.1724 |
| | | SAL | 0.2842 | 0.3006 | 0.2567 | 0.0195 | 0.2510 | 0.8760 | 0.0000 | 0.1741 |
| | | STL | 0.2662 | 0.3236 | 0.2372 | 0.0406 | 0.2441 | 0.8689 | 0.0000 | 0.1759 |
| | 8 | RAN | 0.2773 | 0.3057 | 0.2445 | 0.0252 | 0.2490 | 0.8218 | 0.0000 | 0.1678 |
| | | SAL | 0.2780 | 0.2855 | 0.2718 | 0.0057 | 0.2494 | 0.8686 | 0.0000 | 0.1742 |
| | | STL | 0.2767 | 0.3297 | 0.2427 | 0.0380 | 0.2578 | 0.8571 | 0.0000 | 0.1806 |

Table F.10: grouped-liblinear-twitter-GM2-ALL-ALL-50

| GM2 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | RAN | 0.1731 | 0.2186 | 0.1277 | 0.0455 | 0.1724 | 0.7893 | 0.0000 | 0.1690 |
| | | SAL | 0.1869 | 0.1948 | 0.1789 | 0.0079 | 0.1742 | 0.7926 | 0.0000 | 0.1712 |
| | | STL | 0.1801 | 0.2027 | 0.1576 | 0.0225 | 0.1721 | 0.7589 | 0.0000 | 0.1587 |
| | 1 | RAN | 0.2398 | 0.2568 | 0.2227 | 0.0170 | 0.2098 | 0.8914 | 0.0000 | 0.1693 |
| | | SAL | 0.2410 | 0.2861 | 0.1960 | 0.0451 | 0.2205 | 0.8647 | 0.0000 | 0.1747 |
| | | STL | 0.2242 | 0.2523 | 0.1960 | 0.0282 | 0.2137 | 0.8402 | 0.0000 | 0.1703 |
| | 2 | RAN | 0.2430 | 0.2518 | 0.2341 | 0.0088 | 0.2147 | 0.8397 | 0.0000 | 0.1748 |
| | | SAL | 0.2395 | 0.2697 | 0.2093 | 0.0302 | 0.2119 | 0.8027 | 0.0000 | 0.1694 |
| | | STL | 0.2238 | 0.2609 | 0.1867 | 0.0371 | 0.2124 | 0.8353 | 0.0000 | 0.1688 |
| | 4 | RAN | 0.2424 | 0.2688 | 0.2160 | 0.0264 | 0.2097 | 0.8372 | 0.0000 | 0.1712 |
| | | SAL | 0.2356 | 0.2556 | 0.2156 | 0.0200 | 0.2082 | 0.8615 | 0.0000 | 0.1710 |
| | | STL | 0.2309 | 0.2778 | 0.1841 | 0.0469 | 0.2155 | 0.8291 | 0.0000 | 0.1666 |
| | 8 | RAN | 0.2359 | 0.2575 | 0.2143 | 0.0216 | 0.2135 | 0.8259 | 0.0000 | 0.1699 |
| | | SAL | 0.2445 | 0.2672 | 0.2219 | 0.0227 | 0.2155 | 0.8600 | 0.0000 | 0.1805 |
| | | STL | 0.2257 | 0.2672 | 0.1843 | 0.0415 | 0.2150 | 0.8160 | 0.0000 | 0.1737 |

Table F.11: grouped-liblinear-twitter-GM2-ALL-ALL-75

| GM2 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | RAN | 0.1466 | 0.1466 | 0.1466 | 0.0000 | 0.1383 | 0.7754 | 0.0000 | 0.1565 |
| | | SAL | 0.1554 | 0.1554 | 0.1554 | 0.0000 | 0.1466 | 0.8092 | 0.0000 | 0.1624 |
| | | STL | 0.1554 | 0.1554 | 0.1554 | 0.0000 | 0.1466 | 0.8092 | 0.0000 | 0.1624 |
| | 1 | RAN | 0.1882 | 0.1882 | 0.1882 | 0.0000 | 0.1610 | 0.7742 | 0.0000 | 0.1642 |
| | | SAL | 0.1893 | 0.1893 | 0.1893 | 0.0000 | 0.1683 | 0.7241 | 0.0000 | 0.1618 |
| | | STL | 0.1893 | 0.1893 | 0.1893 | 0.0000 | 0.1683 | 0.7241 | 0.0000 | 0.1618 |
| | 2 | RAN | 0.1874 | 0.1874 | 0.1874 | 0.0000 | 0.1584 | 0.8385 | 0.0000 | 0.1714 |
| | | SAL | 0.1830 | 0.1830 | 0.1830 | 0.0000 | 0.1612 | 0.7421 | 0.0000 | 0.1653 |
| | | STL | 0.1830 | 0.1830 | 0.1830 | 0.0000 | 0.1612 | 0.7421 | 0.0000 | 0.1653 |
| | 4 | RAN | 0.1809 | 0.1809 | 0.1809 | 0.0000 | 0.1573 | 0.7458 | 0.0000 | 0.1599 |
| | | SAL | 0.1888 | 0.1888 | 0.1888 | 0.0000 | 0.1662 | 0.7143 | 0.0000 | 0.1638 |
| | | STL | 0.1888 | 0.1888 | 0.1888 | 0.0000 | 0.1662 | 0.7143 | 0.0000 | 0.1638 |

Table F.12: grouped-liblinear-twitter-GM2-ALL-ALL-150

| GM5 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | RAN | 0.2775 | 0.4753 | 0.1687 | 0.0584 | 0.1980 | 0.6443 | 0.0000 | 0.1541 |
| | | SAL | 0.2754 | 0.5119 | 0.1805 | 0.0606 | 0.1990 | 0.6715 | 0.0000 | 0.1498 |
| | | STL | 0.2763 | 0.4075 | 0.2085 | 0.0483 | 0.2016 | 0.5846 | 0.0000 | 0.1488 |
| | 1 | RAN | 0.5972 | 0.8029 | 0.4040 | 0.0855 | 0.5712 | 0.9498 | 0.1887 | 0.1491 |
| | | SAL | 0.5876 | 0.8362 | 0.4503 | 0.0896 | 0.5650 | 0.9494 | 0.1667 | 0.1503 |
| | | STL | 0.5756 | 0.7406 | 0.4249 | 0.0715 | 0.5609 | 0.9286 | 0.2373 | 0.1405 |
| | 2 | RAN | 0.5836 | 0.7339 | 0.4037 | 0.0852 | 0.5587 | 0.9237 | 0.0845 | 0.1523 |
| | | SAL | 0.5885 | 0.7639 | 0.3689 | 0.0879 | 0.5700 | 0.9457 | 0.1017 | 0.1449 |
| | | STL | 0.5686 | 0.7365 | 0.4480 | 0.0755 | 0.5496 | 0.9105 | 0.2000 | 0.1483 |
| | 4 | SAL | 0.5900 | 0.7832 | 0.4274 | 0.0883 | 0.5684 | 0.9064 | 0.1967 | 0.1439 |
| | | STL | 0.5635 | 0.7406 | 0.3838 | 0.0707 | 0.5481 | 0.9147 | 0.1972 | 0.1421 |
| | | RAN | 0.5905 | 0.7827 | 0.4317 | 0.0894 | 0.5716 | 0.9389 | 0.2295 | 0.1456 |
| | 8 | SAL | 0.5784 | 0.7253 | 0.4162 | 0.0701 | 0.5546 | 0.9506 | 0.1818 | 0.1419 |
| | | STL | 0.5757 | 0.7365 | 0.4249 | 0.0657 | 0.5596 | 0.9105 | 0.2059 | 0.1438 |
| | | RAN | 0.5819 | 0.8039 | 0.4245 | 0.0911 | 0.5643 | 0.9278 | 0.1231 | 0.1463 |
| | 16 | SAL | 0.5800 | 0.7916 | 0.4084 | 0.0966 | 0.5632 | 0.9104 | 0.1765 | 0.1517 |
| | | STL | 0.5661 | 0.7365 | 0.4176 | 0.0710 | 0.5534 | 0.9105 | 0.0625 | 0.1389 |

Table F.13: grouped-liblinear-twitter-GM5-ALL-ALL-5

| GM5 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | RAN | 0.1740 | 0.2808 | 0.1212 | 0.0355 | 0.1302 | 0.5538 | 0.0000 | 0.1202 |
| | | SAL | 0.1727 | 0.2811 | 0.0961 | 0.0439 | 0.1221 | 0.5294 | 0.0000 | 0.1157 |
| | | STL | 0.1688 | 0.2530 | 0.1313 | 0.0304 | 0.1192 | 0.5625 | 0.0000 | 0.1155 |
| | 1 | RAN | 0.4394 | 0.5487 | 0.3428 | 0.0598 | 0.4153 | 0.8931 | 0.0519 | 0.1602 |
| | | SAL | 0.4520 | 0.5825 | 0.3247 | 0.0719 | 0.4245 | 0.9237 | 0.0741 | 0.1673 |
| | | STL | 0.4234 | 0.5439 | 0.3402 | 0.0588 | 0.4053 | 0.8485 | 0.0597 | 0.1544 |
| | 2 | RAN | 0.4501 | 0.5657 | 0.3178 | 0.0665 | 0.4290 | 0.9272 | 0.0000 | 0.1542 |
| | | SAL | 0.4401 | 0.5567 | 0.3453 | 0.0618 | 0.4169 | 0.9302 | 0.0519 | 0.1657 |
| | | STL | 0.4244 | 0.5439 | 0.3108 | 0.0618 | 0.4088 | 0.8640 | 0.0345 | 0.1521 |
| | 4 | RAN | 0.4540 | 0.5758 | 0.3693 | 0.0569 | 0.4296 | 0.9272 | 0.0000 | 0.1579 |
| | | SAL | 0.4499 | 0.5550 | 0.3482 | 0.0490 | 0.4219 | 0.8923 | 0.0000 | 0.1614 |
| | | STL | 0.4198 | 0.5439 | 0.3059 | 0.0623 | 0.4063 | 0.8500 | 0.0328 | 0.1522 |
| | 8 | RAN | 0.4431 | 0.6007 | 0.3683 | 0.0619 | 0.4227 | 0.9219 | 0.1067 | 0.1521 |
| | | SAL | 0.4519 | 0.5438 | 0.3591 | 0.0466 | 0.4204 | 0.9134 | 0.0741 | 0.1611 |
| | | STL | 0.4246 | 0.5439 | 0.3375 | 0.0579 | 0.4122 | 0.8485 | 0.0385 | 0.1483 |
| | 16 | RAN | 0.4564 | 0.5563 | 0.3924 | 0.0554 | 0.4276 | 0.9213 | 0.0370 | 0.1635 |
| | | SAL | 0.4538 | 0.6258 | 0.3294 | 0.0925 | 0.4311 | 0.9160 | 0.0779 | 0.1685 |
| | | STL | 0.4142 | 0.5366 | 0.2951 | 0.0649 | 0.3996 | 0.8321 | 0.0328 | 0.1502 |

Table F.14: grouped-liblinear-twitter-GM5-ALL-ALL-10

| GM5 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | RAN | 0.1053 | 0.1493 | 0.0851 | 0.0211 | 0.0916 | 0.6154 | 0.0000 | 0.1161 |
| | | SAL | 0.1088 | 0.1677 | 0.0717 | 0.0319 | 0.0881 | 0.5075 | 0.0000 | 0.1101 |
| | | STL | 0.1039 | 0.1480 | 0.0816 | 0.0217 | 0.0861 | 0.5758 | 0.0000 | 0.1157 |
| | 1 | RAN | 0.2969 | 0.3466 | 0.2487 | 0.0297 | 0.2722 | 0.8872 | 0.0000 | 0.1590 |
| | | SAL | 0.3133 | 0.3718 | 0.2607 | 0.0398 | 0.2856 | 0.9105 | 0.0000 | 0.1670 |
| | | STL | 0.2932 | 0.3498 | 0.2378 | 0.0332 | 0.2773 | 0.8346 | 0.0000 | 0.1593 |
| | 2 | RAN | 0.3113 | 0.3874 | 0.2285 | 0.0510 | 0.2880 | 0.8913 | 0.0000 | 0.1678 |
| | | SAL | 0.3080 | 0.3700 | 0.2576 | 0.0447 | 0.2838 | 0.8973 | 0.0000 | 0.1609 |
| | | STL | 0.2958 | 0.3352 | 0.1840 | 0.0518 | 0.2801 | 0.8197 | 0.0196 | 0.1655 |
| | 4 | RAN | 0.3019 | 0.3746 | 0.2493 | 0.0436 | 0.2771 | 0.8613 | 0.0000 | 0.1570 |
| | | SAL | 0.3050 | 0.3749 | 0.2491 | 0.0413 | 0.2808 | 0.8627 | 0.0000 | 0.1655 |
| | | STL | 0.2897 | 0.3382 | 0.2324 | 0.0348 | 0.2701 | 0.8217 | 0.0000 | 0.1613 |

Table F.15: grouped-liblinear-twitter-GM5-ALL-ALL-25

| GM5 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | RAN | 0.0827 | 0.1043 | 0.0668 | 0.0158 | 0.0793 | 0.5205 | 0.0000 | 0.1092 |
| | | SAL | 0.0809 | 0.1060 | 0.0666 | 0.0178 | 0.0835 | 0.5333 | 0.0000 | 0.1187 |
| | | STL | 0.0780 | 0.0908 | 0.0678 | 0.0096 | 0.0749 | 0.5373 | 0.0000 | 0.1089 |
| | 1 | RAN | 0.2249 | 0.2511 | 0.1950 | 0.0231 | 0.2079 | 0.8759 | 0.0000 | 0.1559 |
| | | SAL | 0.2266 | 0.2574 | 0.1834 | 0.0315 | 0.2066 | 0.8636 | 0.0000 | 0.1527 |
| | | STL | 0.2220 | 0.2457 | 0.1952 | 0.0207 | 0.2042 | 0.7729 | 0.0000 | 0.1499 |
| | 2 | RAN | 0.2190 | 0.2327 | 0.2035 | 0.0120 | 0.2011 | 0.8803 | 0.0000 | 0.1486 |
| | | SAL | 0.2329 | 0.3069 | 0.1935 | 0.0524 | 0.2146 | 0.8519 | 0.0000 | 0.1633 |
| | | STL | 0.2190 | 0.2430 | 0.1884 | 0.0228 | 0.2020 | 0.7846 | 0.0000 | 0.1528 |

Table F.16: grouped-liblinear-twitter-GM5-ALL-ALL-50

| GM5 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | Group Type | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | RAN | 0.0633 | 0.0681 | 0.0585 | 0.0048 | 0.0625 | 0.4571 | 0.0000 | 0.0971 |
| | | SAL | 0.0607 | 0.0792 | 0.0421 | 0.0186 | 0.0624 | 0.4595 | 0.0000 | 0.0943 |
| | | STL | 0.0690 | 0.0764 | 0.0617 | 0.0073 | 0.0725 | 0.5392 | 0.0000 | 0.1127 |
| | 1 | RAN | 0.1991 | 0.2074 | 0.1907 | 0.0084 | 0.1790 | 0.8326 | 0.0000 | 0.1519 |
| | | SAL | 0.1901 | 0.1943 | 0.1859 | 0.0042 | 0.1714 | 0.8669 | 0.0000 | 0.1535 |
| | | STL | 0.1815 | 0.2133 | 0.1496 | 0.0318 | 0.1668 | 0.7969 | 0.0000 | 0.1486 |
| | 2 | RAN | 0.1944 | 0.2180 | 0.1708 | 0.0236 | 0.1707 | 0.8211 | 0.0000 | 0.1469 |
| | | SAL | 0.1925 | 0.1979 | 0.1870 | 0.0054 | 0.1668 | 0.8201 | 0.0000 | 0.1520 |
| | | STL | 0.1789 | 0.1904 | 0.1673 | 0.0115 | 0.1680 | 0.7333 | 0.0000 | 0.1446 |

Table F.17: grouped-liblinear-twitter-GM5-ALL-ALL-75

| GM5 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | Group Type | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | RAN | 0.0573 | 0.0573 | 0.0573 | 0.0000 | 0.0623 | 0.5363 | 0.0000 | 0.1016 |
| | | SAL | 0.0667 | 0.0667 | 0.0667 | 0.0000 | 0.0749 | 0.5443 | 0.0000 | 0.1102 |
| | | STL | 0.0667 | 0.0667 | 0.0667 | 0.0000 | 0.0749 | 0.5443 | 0.0000 | 0.1102 |
| | 1 | RAN | 0.1402 | 0.1402 | 0.1402 | 0.0000 | 0.1270 | 0.7500 | 0.0000 | 0.1360 |
| | | SAL | 0.1398 | 0.1398 | 0.1398 | 0.0000 | 0.1388 | 0.7399 | 0.0000 | 0.1426 |
| | | STL | 0.1398 | 0.1398 | 0.1398 | 0.0000 | 0.1388 | 0.7399 | 0.0000 | 0.1426 |

Table F.18: grouped-liblinear-twitter-GM5-ALL-ALL-150

| GB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | Group Type | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | RAN | 0.5516 | 0.8313 | 0.3737 | 0.0851 | 0.5171 | 0.9360 | 0.0000 | 0.1613 |
| | | SAL | 0.5413 | 0.7271 | 0.4077 | 0.0823 | 0.5003 | 0.9125 | 0.0000 | 0.1754 |
| | | STL | 0.5286 | 0.7248 | 0.3916 | 0.0846 | 0.5004 | 0.9206 | 0.0656 | 0.1620 |
| | 1 | RAN | 0.5237 | 0.7348 | 0.3734 | 0.0767 | 0.4901 | 0.9308 | 0.0000 | 0.1701 |
| | | SAL | 0.5472 | 0.7360 | 0.3782 | 0.0744 | 0.5070 | 0.9213 | 0.0385 | 0.1749 |
| | | STL | 0.5227 | 0.6789 | 0.3935 | 0.0718 | 0.5016 | 0.9237 | 0.0000 | 0.1594 |
| | 2 | RAN | 0.5529 | 0.6855 | 0.3850 | 0.0687 | 0.5190 | 0.9375 | 0.0800 | 0.1623 |
| | | SAL | 0.5518 | 0.7269 | 0.4286 | 0.0748 | 0.5138 | 0.8897 | 0.0435 | 0.1701 |
| | | STL | 0.5295 | 0.6789 | 0.4142 | 0.0642 | 0.5047 | 0.9231 | 0.0000 | 0.1559 |
| | 4 | RAN | 0.5455 | 0.7530 | 0.3571 | 0.0896 | 0.5146 | 0.9344 | 0.0370 | 0.1655 |
| | | SAL | 0.5573 | 0.7538 | 0.4223 | 0.0816 | 0.5217 | 0.9354 | 0.0435 | 0.1705 |
| | | STL | 0.5167 | 0.6748 | 0.3966 | 0.0730 | 0.4917 | 0.9237 | 0.0000 | 0.1677 |
| | 8 | RAN | 0.5486 | 0.8297 | 0.4236 | 0.0909 | 0.5209 | 0.9290 | 0.0513 | 0.1596 |
| | | SAL | 0.5487 | 0.7169 | 0.3908 | 0.0828 | 0.5056 | 0.8930 | 0.0385 | 0.1772 |
| | | STL | 0.5241 | 0.6789 | 0.3935 | 0.0756 | 0.4989 | 0.9237 | 0.0000 | 0.1616 |
| | 16 | RAN | 0.5473 | 0.7285 | 0.3780 | 0.0773 | 0.5080 | 0.9105 | 0.0000 | 0.1802 |
| | | SAL | 0.5471 | 0.7564 | 0.4084 | 0.0801 | 0.5015 | 0.9416 | 0.0000 | 0.1787 |
| | | STL | 0.5214 | 0.6775 | 0.3966 | 0.0698 | 0.4965 | 0.9237 | 0.0000 | 0.1623 |

Table F.19: grouped-liblinear-twitter-GB3-ALL-ALL-5

| GB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | RAN | 0.4323 | 0.5419 | 0.3306 | 0.0708 | 0.3973 | 0.8931 | 0.0000 | 0.1751 |
| | | SAL | 0.4251 | 0.5716 | 0.3021 | 0.0693 | 0.3946 | 0.9302 | 0.0000 | 0.1727 |
| | | STL | 0.4118 | 0.5448 | 0.3267 | 0.0572 | 0.3819 | 0.9015 | 0.0000 | 0.1714 |
| | 1 | RAN | 0.4203 | 0.5008 | 0.3237 | 0.0515 | 0.3793 | 0.8806 | 0.0000 | 0.1743 |
| | | SAL | 0.4268 | 0.5795 | 0.2998 | 0.0774 | 0.3916 | 0.8655 | 0.0000 | 0.1846 |
| | | STL | 0.4150 | 0.5289 | 0.3373 | 0.0573 | 0.3897 | 0.8727 | 0.0308 | 0.1603 |
| | 2 | RAN | 0.4277 | 0.6907 | 0.3418 | 0.0837 | 0.3982 | 0.9049 | 0.0000 | 0.1809 |
| | | SAL | 0.4239 | 0.5714 | 0.3259 | 0.0686 | 0.3921 | 0.8947 | 0.0385 | 0.1685 |
| | | STL | 0.4146 | 0.5289 | 0.3365 | 0.0600 | 0.3910 | 0.8750 | 0.0000 | 0.1634 |
| | 4 | RAN | 0.4325 | 0.5960 | 0.3114 | 0.0791 | 0.3947 | 0.9127 | 0.0000 | 0.1880 |
| | | SAL | 0.4286 | 0.5818 | 0.3194 | 0.0625 | 0.3922 | 0.9231 | 0.0000 | 0.1746 |
| | | STL | 0.4066 | 0.5280 | 0.3317 | 0.0584 | 0.3827 | 0.8750 | 0.0000 | 0.1635 |
| | 8 | RAN | 0.4295 | 0.5868 | 0.3045 | 0.0738 | 0.3980 | 0.9266 | 0.0000 | 0.1783 |
| | | SAL | 0.4298 | 0.5508 | 0.3273 | 0.0758 | 0.3988 | 0.8947 | 0.0000 | 0.1798 |
| | | STL | 0.4145 | 0.5297 | 0.3274 | 0.0600 | 0.3914 | 0.8750 | 0.0000 | 0.1606 |
| | 16 | RAN | 0.4380 | 0.5621 | 0.3144 | 0.0690 | 0.4078 | 0.9183 | 0.0000 | 0.1714 |
| | | SAL | 0.4291 | 0.5633 | 0.3351 | 0.0633 | 0.3887 | 0.9147 | 0.0000 | 0.1798 |
| | | STL | 0.4043 | 0.5297 | 0.3278 | 0.0622 | 0.3826 | 0.8750 | 0.0000 | 0.1642 |

Table F.20: grouped-liblinear-twitter-GB3-ALL-ALL-10

| GB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | RAN | 0.3201 | 0.3997 | 0.2486 | 0.0509 | 0.2917 | 0.8433 | 0.0000 | 0.1746 |
| | | SAL | 0.3170 | 0.3983 | 0.2658 | 0.0472 | 0.2898 | 0.8750 | 0.0000 | 0.1755 |
| | | STL | 0.2998 | 0.4116 | 0.2218 | 0.0567 | 0.2781 | 0.8531 | 0.0000 | 0.1804 |
| | 1 | RAN | 0.3143 | 0.3864 | 0.2626 | 0.0412 | 0.2831 | 0.9147 | 0.0000 | 0.1686 |
| | | SAL | 0.3172 | 0.3813 | 0.2666 | 0.0401 | 0.2907 | 0.8923 | 0.0000 | 0.1777 |
| | | STL | 0.2946 | 0.4029 | 0.2379 | 0.0547 | 0.2772 | 0.8689 | 0.0000 | 0.1712 |
| | 2 | RAN | 0.3193 | 0.3680 | 0.2569 | 0.0437 | 0.2910 | 0.8949 | 0.0000 | 0.1861 |
| | | SAL | 0.3118 | 0.3603 | 0.2839 | 0.0301 | 0.2706 | 0.8622 | 0.0000 | 0.1778 |
| | | STL | 0.3090 | 0.4094 | 0.2572 | 0.0500 | 0.2872 | 0.8647 | 0.0000 | 0.1696 |
| | 4 | RAN | 0.3191 | 0.3886 | 0.2582 | 0.0424 | 0.2805 | 0.8973 | 0.0000 | 0.1748 |
| | | SAL | 0.3203 | 0.3957 | 0.2714 | 0.0492 | 0.2893 | 0.8841 | 0.0000 | 0.1716 |
| | | STL | 0.3002 | 0.4082 | 0.2461 | 0.0557 | 0.2794 | 0.8593 | 0.0000 | 0.1724 |
| | 8 | RAN | 0.3240 | 0.3914 | 0.2693 | 0.0388 | 0.3031 | 0.9286 | 0.0000 | 0.1813 |
| | | SAL | 0.3225 | 0.3795 | 0.2862 | 0.0348 | 0.2910 | 0.8806 | 0.0000 | 0.1714 |
| | | STL | 0.2982 | 0.4164 | 0.2594 | 0.0554 | 0.2794 | 0.8657 | 0.0000 | 0.1702 |

Table F.21: grouped-liblinear-twitter-GB3-ALL-ALL-25

| GB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | RAN | 0.2506 | 0.2968 | 0.2038 | 0.0379 | 0.2256 | 0.8992 | 0.0000 | 0.1710 |
| | | SAL | 0.2483 | 0.2688 | 0.2372 | 0.0145 | 0.2229 | 0.8540 | 0.0000 | 0.1659 |
| | | STL | 0.2411 | 0.2883 | 0.1960 | 0.0377 | 0.2235 | 0.8298 | 0.0000 | 0.1658 |
| | 1 | RAN | 0.2412 | 0.2821 | 0.2077 | 0.0309 | 0.2161 | 0.8679 | 0.0000 | 0.1683 |
| | | SAL | 0.2572 | 0.2728 | 0.2349 | 0.0161 | 0.2292 | 0.8561 | 0.0000 | 0.1686 |
| | | STL | 0.2412 | 0.3082 | 0.1949 | 0.0485 | 0.2299 | 0.8864 | 0.0000 | 0.1811 |
| | 2 | RAN | 0.2494 | 0.2622 | 0.2352 | 0.0110 | 0.2229 | 0.8812 | 0.0000 | 0.1712 |
| | | SAL | 0.2558 | 0.2942 | 0.2224 | 0.0295 | 0.2295 | 0.8872 | 0.0000 | 0.1661 |
| | | STL | 0.2344 | 0.3034 | 0.1923 | 0.0492 | 0.2205 | 0.9008 | 0.0000 | 0.1717 |
| | 4 | RAN | 0.2532 | 0.2633 | 0.2403 | 0.0096 | 0.2284 | 0.8978 | 0.0000 | 0.1686 |
| | | SAL | 0.2626 | 0.2964 | 0.2126 | 0.0361 | 0.2320 | 0.8699 | 0.0000 | 0.1670 |
| | | STL | 0.2520 | 0.3163 | 0.2019 | 0.0478 | 0.2396 | 0.8764 | 0.0000 | 0.1725 |

Table F.22: grouped-liblinear-twitter-GB3-ALL-ALL-50

| GB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | RAN | 0.2035 | 0.2139 | 0.1931 | 0.0104 | 0.1888 | 0.8496 | 0.0000 | 0.1631 |
| | | SAL | 0.2193 | 0.2325 | 0.2060 | 0.0133 | 0.1972 | 0.8530 | 0.0000 | 0.1658 |
| | | STL | 0.2170 | 0.2497 | 0.1842 | 0.0328 | 0.2029 | 0.8235 | 0.0000 | 0.1732 |
| | 1 | RAN | 0.2184 | 0.2477 | 0.1891 | 0.0293 | 0.1915 | 0.8071 | 0.0000 | 0.1657 |
| | | SAL | 0.2143 | 0.2442 | 0.1845 | 0.0299 | 0.1898 | 0.8803 | 0.0000 | 0.1667 |
| | | STL | 0.2138 | 0.2472 | 0.1803 | 0.0335 | 0.1934 | 0.8561 | 0.0000 | 0.1654 |
| | 2 | RAN | 0.2193 | 0.2448 | 0.1937 | 0.0256 | 0.2010 | 0.8989 | 0.0000 | 0.1717 |
| | | SAL | 0.2243 | 0.2421 | 0.2065 | 0.0178 | 0.1980 | 0.8520 | 0.0000 | 0.1678 |
| | | STL | 0.2085 | 0.2433 | 0.1736 | 0.0348 | 0.1924 | 0.8692 | 0.0000 | 0.1746 |
| | 4 | RAN | 0.2226 | 0.2324 | 0.2128 | 0.0098 | 0.1930 | 0.8846 | 0.0000 | 0.1698 |
| | | SAL | 0.2244 | 0.2648 | 0.1840 | 0.0404 | 0.1947 | 0.8722 | 0.0000 | 0.1672 |
| | | STL | 0.2120 | 0.2428 | 0.1812 | 0.0308 | 0.1985 | 0.8682 | 0.0000 | 0.1696 |

Table F.23: grouped-liblinear-twitter-GB3-ALL-ALL-75

| GB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | Group Type | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | RAN | 0.1825 | 0.1825 | 0.1825 | 0.0000 | 0.1573 | 0.8100 | 0.0000 | 0.1522 |
| | | SAL | 0.1696 | 0.1696 | 0.1696 | 0.0000 | 0.1484 | 0.8357 | 0.0000 | 0.1498 |
| | | STL | 0.1696 | 0.1696 | 0.1696 | 0.0000 | 0.1484 | 0.8357 | 0.0000 | 0.1498 |
| | 1 | RAN | 0.1813 | 0.1813 | 0.1813 | 0.0000 | 0.1581 | 0.8192 | 0.0000 | 0.1611 |
| | | SAL | 0.1796 | 0.1796 | 0.1796 | 0.0000 | 0.1562 | 0.8125 | 0.0000 | 0.1509 |
| | | STL | 0.1796 | 0.1796 | 0.1796 | 0.0000 | 0.1562 | 0.8125 | 0.0000 | 0.1509 |
| | 2 | RAN | 0.1758 | 0.1758 | 0.1758 | 0.0000 | 0.1528 | 0.7723 | 0.0000 | 0.1577 |
| | | SAL | 0.1746 | 0.1746 | 0.1746 | 0.0000 | 0.1568 | 0.8182 | 0.0000 | 0.1558 |
| | | STL | 0.1746 | 0.1746 | 0.1746 | 0.0000 | 0.1568 | 0.8182 | 0.0000 | 0.1558 |

Table F.24: grouped-liblinear-twitter-GB3-ALL-ALL-150

| OSB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | Group Type | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | RAN | 0.5481 | 0.6885 | 0.3680 | 0.0785 | 0.5141 | 0.9272 | 0.0941 | 0.1632 |
| | | SAL | 0.5564 | 0.7559 | 0.3794 | 0.0880 | 0.5289 | 0.9513 | 0.0000 | 0.1634 |
| | | STL | 0.5245 | 0.6680 | 0.3774 | 0.0752 | 0.5001 | 0.9206 | 0.0299 | 0.1644 |
| | 1 | RAN | 0.5480 | 0.7148 | 0.4186 | 0.0746 | 0.5107 | 0.9362 | 0.0952 | 0.1659 |
| | | SAL | 0.5525 | 0.7747 | 0.4000 | 0.0710 | 0.5169 | 0.9290 | 0.0000 | 0.1637 |
| | | STL | 0.5168 | 0.6748 | 0.4059 | 0.0676 | 0.4911 | 0.9237 | 0.0000 | 0.1633 |
| | 2 | RAN | 0.5426 | 0.7221 | 0.3867 | 0.0812 | 0.5010 | 0.9302 | 0.0400 | 0.1739 |
| | | SAL | 0.5623 | 0.7651 | 0.3934 | 0.0786 | 0.5259 | 0.9425 | 0.0000 | 0.1714 |
| | | STL | 0.5233 | 0.6734 | 0.3731 | 0.0720 | 0.4984 | 0.9194 | 0.0000 | 0.1591 |
| | 4 | RAN | 0.5435 | 0.6757 | 0.3967 | 0.0789 | 0.5120 | 0.9434 | 0.0435 | 0.1691 |
| | | SAL | 0.5509 | 0.7747 | 0.3934 | 0.0845 | 0.5152 | 0.9333 | 0.0000 | 0.1704 |
| | | STL | 0.5230 | 0.6762 | 0.4059 | 0.0712 | 0.4983 | 0.9237 | 0.0000 | 0.1619 |
| | 8 | RAN | 0.5520 | 0.7570 | 0.4263 | 0.0827 | 0.5113 | 0.9457 | 0.0000 | 0.1743 |
| | | SAL | 0.5549 | 0.7747 | 0.4000 | 0.0893 | 0.5177 | 0.9425 | 0.0000 | 0.1794 |
| | | STL | 0.5211 | 0.6775 | 0.4143 | 0.0701 | 0.4961 | 0.9280 | 0.0000 | 0.1609 |
| | 16 | RAN | 0.5459 | 0.7647 | 0.3704 | 0.0862 | 0.5088 | 0.9333 | 0.0000 | 0.1626 |
| | | SAL | 0.5557 | 0.7773 | 0.4372 | 0.0780 | 0.5218 | 0.9278 | 0.0000 | 0.1649 |
| | | STL | 0.5200 | 0.6775 | 0.4022 | 0.0665 | 0.4963 | 0.9280 | 0.0000 | 0.1672 |

Table F.25: grouped-liblinear-twitter-OSB3-ALL-ALL-5

| OSB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | RAN | 0.4272 | 0.5191 | 0.3284 | 0.0617 | 0.3974 | 0.9261 | 0.0000 | 0.1711 |
| | | SAL | 0.4408 | 0.5520 | 0.3268 | 0.0575 | 0.4043 | 0.8621 | 0.0000 | 0.1739 |
| | | STL | 0.4135 | 0.5473 | 0.3216 | 0.0532 | 0.3944 | 0.8759 | 0.0426 | 0.1547 |
| | 1 | RAN | 0.4373 | 0.5798 | 0.3328 | 0.0678 | 0.4050 | 0.9453 | 0.0000 | 0.1793 |
| | | SAL | 0.4370 | 0.5847 | 0.3219 | 0.0643 | 0.4003 | 0.8832 | 0.0000 | 0.1769 |
| | | STL | 0.4122 | 0.5331 | 0.3226 | 0.0581 | 0.3864 | 0.8664 | 0.0299 | 0.1756 |
| | 2 | RAN | 0.4246 | 0.5667 | 0.3483 | 0.0546 | 0.3893 | 0.9125 | 0.0000 | 0.1763 |
| | | SAL | 0.4369 | 0.5802 | 0.3380 | 0.0629 | 0.4012 | 0.8812 | 0.0000 | 0.1809 |
| | | STL | 0.4151 | 0.5322 | 0.3280 | 0.0536 | 0.3902 | 0.8633 | 0.0000 | 0.1727 |
| | 4 | RAN | 0.4239 | 0.6093 | 0.3043 | 0.0825 | 0.3891 | 0.9375 | 0.0000 | 0.1804 |
| | | SAL | 0.4318 | 0.6031 | 0.3051 | 0.0734 | 0.3947 | 0.8679 | 0.0000 | 0.1812 |
| | | STL | 0.4193 | 0.5322 | 0.3495 | 0.0479 | 0.3956 | 0.8664 | 0.0392 | 0.1696 |
| | 8 | RAN | 0.4326 | 0.5522 | 0.3042 | 0.0617 | 0.3967 | 0.9358 | 0.0000 | 0.1841 |
| | | SAL | 0.4344 | 0.6031 | 0.3219 | 0.0814 | 0.4011 | 0.9057 | 0.0000 | 0.1818 |
| | | STL | 0.4178 | 0.5322 | 0.3247 | 0.0555 | 0.3939 | 0.8664 | 0.0328 | 0.1734 |

Table F.26: grouped-liblinear-twitter-OSB3-ALL-ALL-10

| OSB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | RAN | 0.3047 | 0.3761 | 0.2584 | 0.0478 | 0.2748 | 0.8722 | 0.0000 | 0.1747 |
| | | SAL | 0.3148 | 0.3923 | 0.2331 | 0.0595 | 0.2924 | 0.8647 | 0.0000 | 0.1755 |
| | | STL | 0.3055 | 0.4086 | 0.2393 | 0.0511 | 0.2876 | 0.8731 | 0.0000 | 0.1790 |
| | 1 | RAN | 0.3188 | 0.4382 | 0.2234 | 0.0653 | 0.2824 | 0.8837 | 0.0000 | 0.1826 |
| | | SAL | 0.3235 | 0.3714 | 0.2589 | 0.0388 | 0.2919 | 0.8945 | 0.0000 | 0.1816 |
| | | STL | 0.3012 | 0.4021 | 0.2604 | 0.0501 | 0.2873 | 0.8571 | 0.0000 | 0.1722 |
| | 2 | RAN | 0.3146 | 0.3747 | 0.2557 | 0.0390 | 0.2824 | 0.9112 | 0.0000 | 0.1774 |
| | | SAL | 0.3142 | 0.3713 | 0.2617 | 0.0436 | 0.2833 | 0.8741 | 0.0000 | 0.1792 |
| | | STL | 0.2898 | 0.4078 | 0.2144 | 0.0593 | 0.2733 | 0.8529 | 0.0000 | 0.1705 |

Table F.27: grouped-liblinear-twitter-OSB3-ALL-ALL-25

| OSB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | RAN | 0.2524 | 0.2630 | 0.2319 | 0.0145 | 0.2311 | 0.8550 | 0.0000 | 0.1770 |
| | | SAL | 0.2589 | 0.2913 | 0.2137 | 0.0329 | 0.2327 | 0.8686 | 0.0000 | 0.1697 |
| | | STL | 0.2446 | 0.2880 | 0.2023 | 0.0350 | 0.2242 | 0.8456 | 0.0000 | 0.1715 |
| | 1 | RAN | 0.2527 | 0.3057 | 0.2138 | 0.0388 | 0.2268 | 0.8727 | 0.0000 | 0.1742 |
| | | SAL | 0.2521 | 0.2849 | 0.2251 | 0.0247 | 0.2250 | 0.8464 | 0.0000 | 0.1739 |
| | | STL | 0.2505 | 0.3164 | 0.2046 | 0.0478 | 0.2323 | 0.8897 | 0.0000 | 0.1766 |

Table F.28: grouped-liblinear-twitter-OSB3-ALL-ALL-50

| OSB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | RAN | 0.2263 | 0.2300 | 0.2227 | 0.0037 | 0.2002 | 0.8374 | 0.0000 | 0.1650 |
| | | SAL | 0.2221 | 0.2493 | 0.1949 | 0.0272 | 0.1998 | 0.8839 | 0.0000 | 0.1735 |
| | | STL | 0.2148 | 0.2482 | 0.1815 | 0.0334 | 0.2028 | 0.8561 | 0.0000 | 0.1740 |
| | 1 | RAN | 0.2217 | 0.2519 | 0.1916 | 0.0302 | 0.1933 | 0.8692 | 0.0000 | 0.1698 |
| | | SAL | 0.2370 | 0.2722 | 0.2019 | 0.0351 | 0.2077 | 0.7818 | 0.0000 | 0.1640 |
| | | STL | 0.2177 | 0.2498 | 0.1856 | 0.0321 | 0.1992 | 0.8303 | 0.0000 | 0.1736 |

Table F.29: grouped-liblinear-twitter-OSB3-ALL-ALL-75

| OSB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | RAN | 0.1839 | 0.1839 | 0.1839 | 0.0000 | 0.1609 | 0.8043 | 0.0000 | 0.1509 |
| | | SAL | 0.1705 | 0.1705 | 0.1705 | 0.0000 | 0.1566 | 0.8239 | 0.0000 | 0.1542 |
| | | STL | 0.1705 | 0.1705 | 0.1705 | 0.0000 | 0.1566 | 0.8239 | 0.0000 | 0.1542 |

Table F.30: grouped-liblinear-twitter-OSB3-ALL-ALL-150

# APPENDIX G:
# Grouped Naive Bayes Results for the ENRON Email Corpus

| GM1 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | Group Type | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | RAN | 0.7112 | 0.8965 | 0.5604 | 0.0942 | 0.3205 | 0.9529 | 0.0000 | 0.3517 |
| | | SAL | 0.6942 | 0.8457 | 0.4841 | 0.0881 | 0.2498 | 0.9165 | 0.0000 | 0.3264 |
| | | STL | 0.7591 | 0.9114 | 0.4815 | 0.0937 | 0.7347 | 0.9730 | 0.0000 | 0.1714 |
| | 1 | RAN | 0.6223 | 0.8864 | 0.2937 | 0.1626 | 0.4969 | 0.9407 | 0.0000 | 0.2549 |
| | | SAL | 0.6009 | 0.8448 | 0.2985 | 0.1576 | 0.4483 | 0.9110 | 0.0000 | 0.2488 |
| | | STL | 0.7091 | 0.8601 | 0.4875 | 0.0791 | 0.6762 | 0.9453 | 0.0000 | 0.1760 |
| | 2 | RAN | 0.6337 | 0.8877 | 0.2256 | 0.1662 | 0.5208 | 0.9420 | 0.0000 | 0.2563 |
| | | SAL | 0.6075 | 0.8704 | 0.2926 | 0.1574 | 0.4534 | 0.9173 | 0.0000 | 0.2476 |
| | | STL | 0.7103 | 0.8612 | 0.5096 | 0.0766 | 0.6787 | 0.9467 | 0.0000 | 0.1729 |
| | 4 | RAN | 0.6545 | 0.8717 | 0.2898 | 0.1469 | 0.5182 | 0.9198 | 0.0000 | 0.2631 |
| | | SAL | 0.6134 | 0.8724 | 0.3059 | 0.1559 | 0.4570 | 0.9186 | 0.0000 | 0.2476 |
| | | STL | 0.7150 | 0.8606 | 0.5176 | 0.0764 | 0.6827 | 0.9483 | 0.0000 | 0.1730 |
| | 8 | RAN | 0.6253 | 0.8864 | 0.2950 | 0.1491 | 0.4916 | 0.9398 | 0.0000 | 0.2540 |
| | | SAL | 0.6184 | 0.8729 | 0.3070 | 0.1541 | 0.4618 | 0.9235 | 0.0000 | 0.2480 |
| | | STL | 0.7167 | 0.8622 | 0.5217 | 0.0756 | 0.6850 | 0.9494 | 0.0000 | 0.1714 |
| | 16 | RAN | 0.6658 | 0.8698 | 0.2551 | 0.1582 | 0.5233 | 0.9173 | 0.0000 | 0.2575 |
| | | SAL | 0.6172 | 0.8599 | 0.2724 | 0.1629 | 0.4627 | 0.9276 | 0.0000 | 0.2480 |
| | | STL | 0.7158 | 0.8628 | 0.5256 | 0.0752 | 0.6840 | 0.9513 | 0.0000 | 0.1715 |

Table G.1: grouped-nb-enron-GM1-ALL-ALL-5

| GM1 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | Group Type | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | RAN | 0.5658 | 0.7384 | 0.3311 | 0.1193 | 0.1854 | 0.8975 | 0.0000 | 0.2631 |
| | | SAL | 0.5240 | 0.7271 | 0.3841 | 0.0955 | 0.1419 | 0.9164 | 0.0000 | 0.2302 |
| | | STL | 0.6406 | 0.7663 | 0.5204 | 0.0710 | 0.6089 | 0.9144 | 0.0000 | 0.1999 |
| | 1 | RAN | 0.4851 | 0.6974 | 0.2923 | 0.1199 | 0.3867 | 0.8844 | 0.0000 | 0.2229 |
| | | SAL | 0.4615 | 0.6259 | 0.3234 | 0.0999 | 0.3605 | 0.8191 | 0.0000 | 0.2192 |
| | | STL | 0.6101 | 0.7117 | 0.3748 | 0.0867 | 0.5791 | 0.9655 | 0.0000 | 0.1960 |
| | 2 | RAN | 0.4861 | 0.7192 | 0.2904 | 0.1308 | 0.3905 | 0.8919 | 0.0000 | 0.2318 |
| | | SAL | 0.4692 | 0.5976 | 0.2952 | 0.0933 | 0.3642 | 0.8037 | 0.0000 | 0.2161 |
| | | STL | 0.6091 | 0.7137 | 0.3918 | 0.0828 | 0.5793 | 0.9157 | 0.0000 | 0.1920 |
| | 4 | RAN | 0.5304 | 0.7545 | 0.2715 | 0.1535 | 0.4113 | 0.8797 | 0.0000 | 0.2443 |
| | | SAL | 0.4767 | 0.6021 | 0.3353 | 0.0915 | 0.3706 | 0.8056 | 0.0000 | 0.2167 |
| | | STL | 0.6145 | 0.7146 | 0.4003 | 0.0822 | 0.5843 | 0.9500 | 0.0000 | 0.1907 |
| | 8 | RAN | 0.5051 | 0.6994 | 0.2647 | 0.1301 | 0.3973 | 0.8365 | 0.0000 | 0.2348 |
| | | SAL | 0.4832 | 0.6110 | 0.3375 | 0.0936 | 0.3763 | 0.8063 | 0.0000 | 0.2187 |
| | | STL | 0.6166 | 0.7164 | 0.4038 | 0.0820 | 0.5864 | 0.9157 | 0.0000 | 0.1897 |
| | 16 | RAN | 0.5172 | 0.6935 | 0.2763 | 0.1387 | 0.3980 | 0.9015 | 0.0000 | 0.2364 |
| | | SAL | 0.4788 | 0.6056 | 0.3219 | 0.0886 | 0.3762 | 0.8341 | 0.0000 | 0.2178 |
| | | STL | 0.6170 | 0.7174 | 0.4079 | 0.0812 | 0.5869 | 0.9500 | 0.0000 | 0.1899 |

Table G.2: grouped-nb-enron-GM1-ALL-ALL-10

| GM1 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | Group Type | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | RAN | 0.4141 | 0.5192 | 0.3471 | 0.0548 | 0.1008 | 0.8796 | 0.0000 | 0.1952 |
| | | SAL | 0.3826 | 0.4585 | 0.2996 | 0.0622 | 0.0896 | 0.8666 | 0.0000 | 0.1791 |
| | | STL | 0.4280 | 0.5172 | 0.3652 | 0.0466 | 0.3652 | 0.8750 | 0.0000 | 0.2397 |
| | 1 | RAN | 0.3598 | 0.4587 | 0.2930 | 0.0601 | 0.2919 | 0.8148 | 0.0000 | 0.2012 |
| | | SAL | 0.3306 | 0.3737 | 0.2745 | 0.0367 | 0.2805 | 0.8205 | 0.0000 | 0.1863 |
| | | STL | 0.4964 | 0.5915 | 0.3156 | 0.0916 | 0.4648 | 0.9870 | 0.0000 | 0.2020 |
| | 2 | RAN | 0.3504 | 0.4163 | 0.3035 | 0.0373 | 0.2879 | 0.7757 | 0.0000 | 0.1908 |
| | | SAL | 0.3618 | 0.4662 | 0.2822 | 0.0656 | 0.2896 | 0.7686 | 0.0000 | 0.1969 |
| | | STL | 0.4989 | 0.5964 | 0.3246 | 0.0906 | 0.4688 | 0.9870 | 0.0000 | 0.2005 |
| | 4 | RAN | 0.3616 | 0.4911 | 0.2402 | 0.0871 | 0.3053 | 0.8243 | 0.0000 | 0.2001 |
| | | SAL | 0.3690 | 0.4729 | 0.2850 | 0.0654 | 0.2957 | 0.7718 | 0.0000 | 0.1986 |
| | | STL | 0.5026 | 0.5966 | 0.3294 | 0.0883 | 0.4739 | 0.9870 | 0.0000 | 0.1990 |
| | 8 | RAN | 0.3580 | 0.4243 | 0.2544 | 0.0568 | 0.3017 | 0.8513 | 0.0000 | 0.1990 |
| | | SAL | 0.3758 | 0.4772 | 0.3050 | 0.0630 | 0.2989 | 0.7813 | 0.0000 | 0.1991 |
| | | STL | 0.5041 | 0.5982 | 0.3330 | 0.0871 | 0.4753 | 0.9870 | 0.0000 | 0.1979 |
| | 16 | RAN | 0.3729 | 0.4546 | 0.3235 | 0.0407 | 0.3039 | 0.8175 | 0.0000 | 0.1978 |
| | | SAL | 0.3715 | 0.4700 | 0.2909 | 0.0603 | 0.2998 | 0.7713 | 0.0000 | 0.1978 |
| | | STL | 0.5055 | 0.5986 | 0.3372 | 0.0857 | 0.4761 | 0.9870 | 0.0000 | 0.1981 |

Table G.3: grouped-nb-enron-GM1-ALL-ALL-25

| GM1 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | RAN | 0.3205 | 0.4130 | 0.2686 | 0.0656 | 0.0672 | 0.8599 | 0.0000 | 0.1603 |
| | | SAL | 0.3237 | 0.3750 | 0.2792 | 0.0394 | 0.0685 | 0.8667 | 0.0000 | 0.1647 |
| | | STL | 0.2936 | 0.2978 | 0.2901 | 0.0032 | 0.1921 | 0.8718 | 0.0000 | 0.2142 |
| | 1 | RAN | 0.2744 | 0.3331 | 0.2307 | 0.0431 | 0.2497 | 0.8040 | 0.0000 | 0.1828 |
| | | SAL | 0.2623 | 0.2763 | 0.2495 | 0.0110 | 0.2404 | 0.7367 | 0.0000 | 0.1724 |
| | | STL | 0.4092 | 0.4779 | 0.2728 | 0.0965 | 0.3853 | 0.9157 | 0.0000 | 0.2012 |
| | 2 | RAN | 0.2681 | 0.2885 | 0.2549 | 0.0146 | 0.2486 | 0.7320 | 0.0000 | 0.1797 |
| | | SAL | 0.2749 | 0.2802 | 0.2681 | 0.0051 | 0.2459 | 0.7417 | 0.0000 | 0.1773 |
| | | STL | 0.4144 | 0.4838 | 0.2781 | 0.0964 | 0.3904 | 0.9157 | 0.0000 | 0.2032 |
| | 4 | RAN | 0.2862 | 0.3415 | 0.2526 | 0.0394 | 0.2552 | 0.6896 | 0.0000 | 0.1749 |
| | | SAL | 0.2922 | 0.3067 | 0.2778 | 0.0118 | 0.2533 | 0.7423 | 0.0000 | 0.1804 |
| | | STL | 0.4176 | 0.4842 | 0.2850 | 0.0937 | 0.3949 | 0.9157 | 0.0000 | 0.2014 |
| | 8 | RAN | 0.2819 | 0.3057 | 0.2641 | 0.0175 | 0.2593 | 0.7625 | 0.0000 | 0.1854 |
| | | SAL | 0.2865 | 0.2916 | 0.2824 | 0.0038 | 0.2524 | 0.7513 | 0.0000 | 0.1779 |
| | | STL | 0.4203 | 0.4864 | 0.2885 | 0.0932 | 0.3973 | 0.9157 | 0.0000 | 0.2013 |
| | 16 | RAN | 0.2952 | 0.3505 | 0.2587 | 0.0398 | 0.2585 | 0.7603 | 0.0000 | 0.1842 |
| | | SAL | 0.3001 | 0.3139 | 0.2898 | 0.0101 | 0.2596 | 0.7871 | 0.0000 | 0.1854 |
| | | STL | 0.4218 | 0.4875 | 0.2919 | 0.0919 | 0.3975 | 0.9157 | 0.0000 | 0.2015 |

Table G.4: grouped-nb-enron-GM1-ALL-ALL-50

| GM1 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | Group Type | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | RAN | 0.3069 | 0.3441 | 0.2696 | 0.0373 | 0.0618 | 0.8578 | 0.0000 | 0.1515 |
| | | SAL | 0.2796 | 0.2988 | 0.2603 | 0.0193 | 0.0579 | 0.8647 | 0.0000 | 0.1534 |
| | | STL | 0.2871 | 0.3076 | 0.2667 | 0.0204 | 0.1177 | 0.8705 | 0.0000 | 0.1836 |
| | 1 | RAN | 0.2345 | 0.2603 | 0.2087 | 0.0258 | 0.2265 | 0.7585 | 0.0000 | 0.1717 |
| | | SAL | 0.2287 | 0.2468 | 0.2106 | 0.0181 | 0.2236 | 0.7277 | 0.0000 | 0.1686 |
| | | STL | 0.3250 | 0.4048 | 0.2451 | 0.0798 | 0.3197 | 0.8085 | 0.0000 | 0.1872 |
| | 2 | RAN | 0.2518 | 0.3025 | 0.2011 | 0.0507 | 0.2314 | 0.7606 | 0.0000 | 0.1766 |
| | | SAL | 0.2589 | 0.2968 | 0.2210 | 0.0379 | 0.2328 | 0.6919 | 0.0000 | 0.1773 |
| | | STL | 0.3293 | 0.4078 | 0.2509 | 0.0785 | 0.3235 | 0.8172 | 0.0000 | 0.1888 |
| | 4 | RAN | 0.2410 | 0.2438 | 0.2382 | 0.0028 | 0.2283 | 0.7446 | 0.0000 | 0.1714 |
| | | SAL | 0.2535 | 0.2814 | 0.2256 | 0.0279 | 0.2354 | 0.7226 | 0.0000 | 0.1774 |
| | | STL | 0.3337 | 0.4106 | 0.2569 | 0.0768 | 0.3285 | 0.8000 | 0.0000 | 0.1880 |
| | 8 | RAN | 0.2462 | 0.2688 | 0.2237 | 0.0225 | 0.2343 | 0.6904 | 0.0000 | 0.1723 |
| | | SAL | 0.2678 | 0.3051 | 0.2305 | 0.0373 | 0.2393 | 0.7048 | 0.0000 | 0.1788 |
| | | STL | 0.3359 | 0.4115 | 0.2604 | 0.0755 | 0.3295 | 0.8261 | 0.0000 | 0.1896 |
| | 16 | RAN | 0.2676 | 0.2869 | 0.2483 | 0.0193 | 0.2396 | 0.7419 | 0.0000 | 0.1833 |
| | | SAL | 0.2588 | 0.2895 | 0.2282 | 0.0307 | 0.2381 | 0.7295 | 0.0000 | 0.1771 |
| | | STL | 0.3388 | 0.4136 | 0.2639 | 0.0748 | 0.3319 | 0.7917 | 0.0000 | 0.1901 |

Table G.5: grouped-nb-enron-GM1-ALL-ALL-75

| GM1 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | Group Type | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | RAN | 0.2450 | 0.2450 | 0.2450 | 0.0000 | 0.0487 | 0.8668 | 0.0000 | 0.1401 |
| | | SAL | 0.2451 | 0.2451 | 0.2451 | 0.0000 | 0.0488 | 0.8674 | 0.0000 | 0.1403 |
| | | STL | 0.2451 | 0.2451 | 0.2451 | 0.0000 | 0.0488 | 0.8674 | 0.0000 | 0.1403 |
| | 1 | RAN | 0.1839 | 0.1839 | 0.1839 | 0.0000 | 0.1942 | 0.6636 | 0.0000 | 0.1582 |
| | | SAL | 0.1841 | 0.1841 | 0.1841 | 0.0000 | 0.1935 | 0.6728 | 0.0000 | 0.1572 |
| | | STL | 0.1841 | 0.1841 | 0.1841 | 0.0000 | 0.1935 | 0.6728 | 0.0000 | 0.1572 |
| | 2 | RAN | 0.1893 | 0.1893 | 0.1893 | 0.0000 | 0.1965 | 0.6773 | 0.0000 | 0.1592 |
| | | SAL | 0.1901 | 0.1901 | 0.1901 | 0.0000 | 0.1974 | 0.6735 | 0.0000 | 0.1591 |
| | | STL | 0.1901 | 0.1901 | 0.1901 | 0.0000 | 0.1974 | 0.6735 | 0.0000 | 0.1591 |
| | 4 | RAN | 0.1956 | 0.1956 | 0.1956 | 0.0000 | 0.2016 | 0.6844 | 0.0000 | 0.1612 |
| | | SAL | 0.1955 | 0.1955 | 0.1955 | 0.0000 | 0.2016 | 0.6801 | 0.0000 | 0.1600 |
| | | STL | 0.1955 | 0.1955 | 0.1955 | 0.0000 | 0.2016 | 0.6801 | 0.0000 | 0.1600 |
| | 8 | RAN | 0.1989 | 0.1989 | 0.1989 | 0.0000 | 0.2031 | 0.6986 | 0.0000 | 0.1618 |
| | | SAL | 0.1991 | 0.1991 | 0.1991 | 0.0000 | 0.2036 | 0.6801 | 0.0000 | 0.1615 |
| | | STL | 0.1991 | 0.1991 | 0.1991 | 0.0000 | 0.2036 | 0.6801 | 0.0000 | 0.1615 |
| | 16 | RAN | 0.2018 | 0.2018 | 0.2018 | 0.0000 | 0.2052 | 0.6794 | 0.0000 | 0.1622 |
| | | SAL | 0.2028 | 0.2028 | 0.2028 | 0.0000 | 0.2056 | 0.6926 | 0.0000 | 0.1630 |
| | | STL | 0.2028 | 0.2028 | 0.2028 | 0.0000 | 0.2056 | 0.6926 | 0.0000 | 0.1630 |

Table G.6: grouped-nb-enron-GM1-ALL-ALL-150

| GM2 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | RAN | 0.8114 | 0.9270 | 0.6573 | 0.0662 | 0.4915 | 0.9781 | 0.0000 | 0.3374 |
| | | SAL | 0.7950 | 0.9010 | 0.6311 | 0.0598 | 0.4441 | 0.9555 | 0.0000 | 0.3345 |
| | | STL | 0.8119 | 0.9337 | 0.5185 | 0.0910 | 0.7841 | 0.9670 | 0.0000 | 0.1661 |
| | 1 | RAN | 0.6333 | 0.8675 | 0.2766 | 0.1733 | 0.5020 | 0.9370 | 0.0000 | 0.2578 |
| | | SAL | 0.6308 | 0.8528 | 0.2883 | 0.1638 | 0.4829 | 0.9281 | 0.0000 | 0.2506 |
| | | STL | 0.6968 | 0.8763 | 0.5445 | 0.0787 | 0.6739 | 0.9529 | 0.0000 | 0.1532 |
| | 2 | RAN | 0.7264 | 0.9132 | 0.4417 | 0.1062 | 0.5780 | 0.9545 | 0.0000 | 0.2582 |
| | | SAL | 0.6882 | 0.8550 | 0.3035 | 0.1284 | 0.5210 | 0.9321 | 0.0000 | 0.2516 |
| | | STL | 0.7186 | 0.8705 | 0.6072 | 0.0766 | 0.7004 | 1.0000 | 0.0000 | 0.1422 |
| | 4 | RAN | 0.7356 | 0.8675 | 0.4845 | 0.1042 | 0.5817 | 0.9253 | 0.0000 | 0.2450 |
| | | SAL | 0.7410 | 0.8782 | 0.4797 | 0.1032 | 0.5602 | 0.9559 | 0.0000 | 0.2553 |
| | | STL | 0.7194 | 0.8899 | 0.5837 | 0.0757 | 0.6987 | 0.9656 | 0.0000 | 0.1439 |
| | 8 | RAN | 0.8166 | 0.9224 | 0.6607 | 0.0680 | 0.6420 | 0.9642 | 0.0000 | 0.2402 |
| | | SAL | 0.8242 | 0.9216 | 0.5879 | 0.0723 | 0.6286 | 0.9755 | 0.0000 | 0.2477 |
| | | STL | 0.7474 | 0.8784 | 0.6321 | 0.0597 | 0.7304 | 0.9630 | 0.0000 | 0.1242 |
| | 16 | RAN | 0.8333 | 0.9489 | 0.6951 | 0.0657 | 0.6524 | 0.9759 | 0.0000 | 0.2473 |
| | | SAL | 0.8572 | 0.9347 | 0.7479 | 0.0455 | 0.6354 | 0.9746 | 0.0000 | 0.2634 |
| | | STL | 0.7571 | 0.8821 | 0.5926 | 0.0649 | 0.7377 | 0.9630 | 0.0000 | 0.1422 |

Table G.7: grouped-nb-enron-GM2-ALL-ALL-5

| GM2 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | RAN | 0.7271 | 0.8713 | 0.5381 | 0.0877 | 0.4106 | 0.9710 | 0.0000 | 0.3276 |
| | | SAL | 0.6963 | 0.8205 | 0.5599 | 0.0675 | 0.3439 | 0.9394 | 0.0000 | 0.3055 |
| | | STL | 0.7393 | 0.9024 | 0.5745 | 0.0901 | 0.7041 | 0.9679 | 0.0000 | 0.1959 |
| | 1 | RAN | 0.5131 | 0.7902 | 0.2541 | 0.1471 | 0.4019 | 0.8853 | 0.0000 | 0.2319 |
| | | SAL | 0.4976 | 0.6377 | 0.3167 | 0.1058 | 0.3901 | 0.8932 | 0.0000 | 0.2245 |
| | | STL | 0.6089 | 0.6987 | 0.4849 | 0.0610 | 0.5794 | 0.8749 | 0.0000 | 0.1824 |
| | 2 | RAN | 0.5793 | 0.7090 | 0.3424 | 0.1013 | 0.4527 | 0.8739 | 0.0000 | 0.2388 |
| | | SAL | 0.5539 | 0.7023 | 0.3271 | 0.1131 | 0.4278 | 0.8901 | 0.0000 | 0.2287 |
| | | STL | 0.6209 | 0.7330 | 0.5128 | 0.0545 | 0.5951 | 0.9655 | 0.0000 | 0.1823 |
| | 4 | RAN | 0.6270 | 0.8022 | 0.4944 | 0.0828 | 0.4931 | 0.9196 | 0.0000 | 0.2395 |
| | | SAL | 0.6081 | 0.7190 | 0.4823 | 0.0745 | 0.4607 | 0.9005 | 0.0000 | 0.2312 |
| | | STL | 0.6302 | 0.7094 | 0.5207 | 0.0585 | 0.5991 | 0.9241 | 0.0000 | 0.1836 |
| | 8 | RAN | 0.7474 | 0.8440 | 0.6015 | 0.0625 | 0.5565 | 0.9400 | 0.0000 | 0.2302 |
| | | SAL | 0.7311 | 0.8202 | 0.5846 | 0.0722 | 0.5474 | 0.9410 | 0.0000 | 0.2332 |
| | | STL | 0.6605 | 0.7617 | 0.5489 | 0.0532 | 0.6345 | 0.9286 | 0.0000 | 0.1705 |
| | 16 | RAN | 0.7656 | 0.8961 | 0.5957 | 0.0807 | 0.5709 | 0.9759 | 0.0000 | 0.2469 |
| | | SAL | 0.7818 | 0.8610 | 0.6899 | 0.0460 | 0.5656 | 0.9471 | 0.0000 | 0.2546 |
| | | STL | 0.6728 | 0.7552 | 0.5735 | 0.0520 | 0.6487 | 0.9500 | 0.0000 | 0.1723 |

Table G.8: grouped-nb-enron-GM2-ALL-ALL-10

| GM2 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | RAN | 0.5975 | 0.7145 | 0.4636 | 0.0953 | 0.2695 | 0.9293 | 0.0000 | 0.2761 |
| | | SAL | 0.5942 | 0.7023 | 0.4523 | 0.0743 | 0.2636 | 0.9152 | 0.0000 | 0.2805 |
| | | STL | 0.6331 | 0.7088 | 0.5554 | 0.0571 | 0.5768 | 0.9737 | 0.0000 | 0.2225 |
| | 1 | RAN | 0.3720 | 0.4786 | 0.2837 | 0.0763 | 0.3161 | 0.8705 | 0.0000 | 0.2091 |
| | | SAL | 0.3766 | 0.4937 | 0.2742 | 0.0667 | 0.3146 | 0.8975 | 0.0000 | 0.2092 |
| | | STL | 0.5011 | 0.5881 | 0.4209 | 0.0580 | 0.4717 | 0.8636 | 0.0000 | 0.1968 |
| | 2 | RAN | 0.4446 | 0.5200 | 0.3110 | 0.0839 | 0.3586 | 0.8620 | 0.0000 | 0.2170 |
| | | SAL | 0.4214 | 0.4976 | 0.3305 | 0.0498 | 0.3458 | 0.8644 | 0.0000 | 0.2116 |
| | | STL | 0.5153 | 0.6078 | 0.4446 | 0.0534 | 0.4907 | 0.9231 | 0.0000 | 0.1869 |
| | 4 | RAN | 0.4960 | 0.5403 | 0.4444 | 0.0348 | 0.3808 | 0.8547 | 0.0000 | 0.2260 |
| | | SAL | 0.4878 | 0.5873 | 0.4101 | 0.0570 | 0.3748 | 0.9188 | 0.0000 | 0.2253 |
| | | STL | 0.5207 | 0.5802 | 0.4948 | 0.0302 | 0.4916 | 0.8941 | 0.0000 | 0.1864 |
| | 8 | RAN | 0.6310 | 0.7008 | 0.5261 | 0.0557 | 0.4663 | 0.9257 | 0.0000 | 0.2214 |
| | | SAL | 0.6211 | 0.7139 | 0.5706 | 0.0442 | 0.4560 | 0.8951 | 0.0000 | 0.2196 |
| | | STL | 0.5605 | 0.6335 | 0.5254 | 0.0355 | 0.5331 | 0.9009 | 0.0000 | 0.1806 |
| | 16 | RAN | 0.6782 | 0.7517 | 0.6019 | 0.0625 | 0.4939 | 0.9191 | 0.0000 | 0.2378 |
| | | SAL | 0.6826 | 0.7904 | 0.5595 | 0.0833 | 0.4880 | 0.9275 | 0.0000 | 0.2514 |
| | | STL | 0.5799 | 0.6454 | 0.5391 | 0.0428 | 0.5538 | 0.9867 | 0.0000 | 0.1885 |

Table G.9: grouped-nb-enron-GM2-ALL-ALL-25

| GM2 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | RAN | 0.5627 | 0.5970 | 0.5242 | 0.0299 | 0.2547 | 0.9063 | 0.0000 | 0.2697 |
| | | SAL | 0.5338 | 0.5932 | 0.4515 | 0.0601 | 0.2180 | 0.8959 | 0.0000 | 0.2590 |
| | | STL | 0.5278 | 0.5459 | 0.4946 | 0.0236 | 0.4315 | 0.9296 | 0.0000 | 0.2511 |
| | 1 | RAN | 0.3025 | 0.3280 | 0.2742 | 0.0220 | 0.2681 | 0.7377 | 0.0000 | 0.1933 |
| | | SAL | 0.3090 | 0.3163 | 0.3044 | 0.0052 | 0.2728 | 0.8958 | 0.0000 | 0.2016 |
| | | STL | 0.4228 | 0.4884 | 0.3114 | 0.0792 | 0.3945 | 0.8132 | 0.0000 | 0.2012 |
| | 2 | RAN | 0.3539 | 0.3809 | 0.3151 | 0.0281 | 0.2984 | 0.8045 | 0.0000 | 0.1988 |
| | | SAL | 0.3547 | 0.3661 | 0.3448 | 0.0088 | 0.2971 | 0.8347 | 0.0000 | 0.1989 |
| | | STL | 0.4408 | 0.5023 | 0.3303 | 0.0783 | 0.4158 | 0.8128 | 0.0000 | 0.1979 |
| | 4 | RAN | 0.4102 | 0.4249 | 0.3939 | 0.0127 | 0.3252 | 0.8479 | 0.0000 | 0.2087 |
| | | SAL | 0.4134 | 0.4192 | 0.4044 | 0.0064 | 0.3268 | 0.8974 | 0.0000 | 0.2160 |
| | | STL | 0.4542 | 0.4894 | 0.3918 | 0.0442 | 0.4233 | 0.8176 | 0.0000 | 0.1953 |
| | 8 | RAN | 0.5480 | 0.5908 | 0.5177 | 0.0311 | 0.4041 | 0.8840 | 0.0000 | 0.2085 |
| | | SAL | 0.5554 | 0.5911 | 0.5344 | 0.0253 | 0.4017 | 0.8618 | 0.0000 | 0.2118 |
| | | STL | 0.5124 | 0.5370 | 0.4971 | 0.0175 | 0.4688 | 0.8941 | 0.0000 | 0.1896 |
| | 16 | RAN | 0.6202 | 0.6477 | 0.5704 | 0.0353 | 0.4438 | 0.9051 | 0.0000 | 0.2349 |
| | | SAL | 0.6169 | 0.6972 | 0.5731 | 0.0568 | 0.4410 | 0.9356 | 0.0000 | 0.2435 |
| | | STL | 0.5301 | 0.5543 | 0.4888 | 0.0294 | 0.4883 | 0.9589 | 0.0000 | 0.2016 |

Table G.10: grouped-nb-enron-GM2-ALL-ALL-50

| GM2 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | RAN | 0.5153 | 0.5296 | 0.5010 | 0.0143 | 0.2140 | 0.8921 | 0.0000 | 0.2510 |
| | | SAL | 0.5030 | 0.5128 | 0.4932 | 0.0098 | 0.1989 | 0.8916 | 0.0000 | 0.2485 |
| | | STL | 0.4987 | 0.5048 | 0.4925 | 0.0062 | 0.3328 | 0.8700 | 0.0000 | 0.2604 |
| | 1 | RAN | 0.2635 | 0.2643 | 0.2626 | 0.0008 | 0.2461 | 0.6999 | 0.0000 | 0.1890 |
| | | SAL | 0.2678 | 0.3060 | 0.2295 | 0.0382 | 0.2480 | 0.7862 | 0.0000 | 0.1931 |
| | | STL | 0.3375 | 0.4085 | 0.2664 | 0.0711 | 0.3161 | 0.8282 | 0.0000 | 0.1847 |
| | 2 | RAN | 0.3145 | 0.3434 | 0.2857 | 0.0289 | 0.2789 | 0.7465 | 0.0000 | 0.1964 |
| | | SAL | 0.3092 | 0.3473 | 0.2711 | 0.0381 | 0.2752 | 0.8037 | 0.0000 | 0.1951 |
| | | STL | 0.3620 | 0.4265 | 0.2974 | 0.0645 | 0.3514 | 0.8235 | 0.0000 | 0.1879 |
| | 4 | RAN | 0.3731 | 0.4061 | 0.3401 | 0.0330 | 0.3039 | 0.8706 | 0.0000 | 0.2099 |
| | | SAL | 0.3657 | 0.4049 | 0.3264 | 0.0392 | 0.3012 | 0.8670 | 0.0000 | 0.2100 |
| | | STL | 0.3900 | 0.4361 | 0.3439 | 0.0461 | 0.3687 | 0.8291 | 0.0000 | 0.1892 |
| | 8 | RAN | 0.5275 | 0.5650 | 0.4901 | 0.0374 | 0.3768 | 0.8579 | 0.0000 | 0.2108 |
| | | SAL | 0.5069 | 0.5458 | 0.4681 | 0.0388 | 0.3722 | 0.8737 | 0.0000 | 0.2070 |
| | | STL | 0.4710 | 0.4795 | 0.4625 | 0.0085 | 0.4212 | 0.8429 | 0.0000 | 0.1836 |
| | 16 | RAN | 0.5897 | 0.6703 | 0.5092 | 0.0806 | 0.4142 | 0.9144 | 0.0000 | 0.2394 |
| | | SAL | 0.5758 | 0.6406 | 0.5111 | 0.0647 | 0.4100 | 0.8824 | 0.0000 | 0.2412 |
| | | STL | 0.4985 | 0.5315 | 0.4654 | 0.0331 | 0.4475 | 0.9143 | 0.0000 | 0.2082 |

Table G.11: grouped-nb-enron-GM2-ALL-ALL-75

| GM2 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | RAN | 0.4535 | 0.4535 | 0.4535 | 0.0000 | 0.1703 | 0.8555 | 0.0000 | 0.2304 |
| | | SAL | 0.4537 | 0.4537 | 0.4537 | 0.0000 | 0.1708 | 0.8573 | 0.0000 | 0.2301 |
| | | STL | 0.4537 | 0.4537 | 0.4537 | 0.0000 | 0.1708 | 0.8573 | 0.0000 | 0.2301 |
| | 1 | RAN | 0.2163 | 0.2163 | 0.2163 | 0.0000 | 0.2158 | 0.6874 | 0.0000 | 0.1767 |
| | | SAL | 0.2164 | 0.2164 | 0.2164 | 0.0000 | 0.2160 | 0.6834 | 0.0000 | 0.1780 |
| | | STL | 0.2164 | 0.2164 | 0.2164 | 0.0000 | 0.2160 | 0.6834 | 0.0000 | 0.1780 |
| | 2 | RAN | 0.2593 | 0.2593 | 0.2593 | 0.0000 | 0.2403 | 0.7671 | 0.0000 | 0.1822 |
| | | SAL | 0.2601 | 0.2601 | 0.2601 | 0.0000 | 0.2403 | 0.7682 | 0.0000 | 0.1824 |
| | | STL | 0.2601 | 0.2601 | 0.2601 | 0.0000 | 0.2403 | 0.7682 | 0.0000 | 0.1824 |
| | 4 | RAN | 0.3097 | 0.3097 | 0.3097 | 0.0000 | 0.2657 | 0.8364 | 0.0000 | 0.1964 |
| | | SAL | 0.3095 | 0.3095 | 0.3095 | 0.0000 | 0.2660 | 0.8385 | 0.0000 | 0.1972 |
| | | STL | 0.3095 | 0.3095 | 0.3095 | 0.0000 | 0.2660 | 0.8385 | 0.0000 | 0.1972 |
| | 8 | RAN | 0.4539 | 0.4539 | 0.4539 | 0.0000 | 0.3333 | 0.8287 | 0.0000 | 0.2013 |
| | | SAL | 0.4552 | 0.4552 | 0.4552 | 0.0000 | 0.3333 | 0.8334 | 0.0000 | 0.2011 |
| | | STL | 0.4552 | 0.4552 | 0.4552 | 0.0000 | 0.3333 | 0.8334 | 0.0000 | 0.2011 |
| | 16 | RAN | 0.5058 | 0.5058 | 0.5058 | 0.0000 | 0.3738 | 0.8657 | 0.0000 | 0.2349 |
| | | SAL | 0.5063 | 0.5063 | 0.5063 | 0.0000 | 0.3733 | 0.8657 | 0.0000 | 0.2352 |
| | | STL | 0.5063 | 0.5063 | 0.5063 | 0.0000 | 0.3733 | 0.8657 | 0.0000 | 0.2352 |

Table G.12: grouped-nb-enron-GM2-ALL-ALL-150

| GM5 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | RAN | 0.7692 | 0.9618 | 0.6018 | 0.1082 | 0.5340 | 0.9802 | 0.0000 | 0.3121 |
| | | SAL | 0.8106 | 0.9055 | 0.6539 | 0.0704 | 0.5305 | 0.9528 | 0.0000 | 0.3241 |
| | | STL | 0.6339 | 0.8767 | 0.4180 | 0.1223 | 0.5810 | 0.9870 | 0.0000 | 0.2393 |
| | 1 | RAN | 0.7957 | 0.9380 | 0.6322 | 0.0867 | 0.6377 | 0.9693 | 0.0000 | 0.2332 |
| | | SAL | 0.8073 | 0.9197 | 0.5353 | 0.0788 | 0.6169 | 0.9586 | 0.0000 | 0.2497 |
| | | STL | 0.7421 | 0.8766 | 0.6145 | 0.0651 | 0.7247 | 0.9600 | 0.0000 | 0.1426 |
| | 2 | RAN | 0.8289 | 0.9554 | 0.6859 | 0.0629 | 0.6597 | 0.9806 | 0.0000 | 0.2434 |
| | | SAL | 0.8519 | 0.9318 | 0.7022 | 0.0562 | 0.6470 | 1.0000 | 0.0000 | 0.2618 |
| | | STL | 0.7504 | 0.9077 | 0.6325 | 0.0657 | 0.7327 | 0.9540 | 0.0000 | 0.1362 |
| | 4 | RAN | 0.8427 | 0.9428 | 0.6772 | 0.0533 | 0.6313 | 0.9698 | 0.0000 | 0.2783 |
| | | SAL | 0.8553 | 0.9436 | 0.7293 | 0.0428 | 0.6143 | 0.9693 | 0.0000 | 0.2859 |
| | | STL | 0.7638 | 0.9006 | 0.6265 | 0.0668 | 0.7478 | 0.9500 | 0.0000 | 0.1351 |
| | 8 | RAN | 0.8120 | 0.9372 | 0.6667 | 0.0776 | 0.5955 | 0.9718 | 0.0000 | 0.2812 |
| | | SAL | 0.8325 | 0.9241 | 0.7240 | 0.0472 | 0.5585 | 0.9582 | 0.0000 | 0.2960 |
| | | STL | 0.7747 | 0.9218 | 0.6265 | 0.0740 | 0.7590 | 0.9559 | 0.0000 | 0.1401 |
| | 16 | RAN | 0.8068 | 0.9380 | 0.6991 | 0.0625 | 0.5363 | 0.9676 | 0.0000 | 0.3012 |
| | | SAL | 0.8126 | 0.9073 | 0.6810 | 0.0567 | 0.5136 | 0.9494 | 0.0000 | 0.3063 |
| | | STL | 0.7746 | 0.9065 | 0.6325 | 0.0718 | 0.7596 | 0.9557 | 0.0000 | 0.1391 |

Table G.13: grouped-nb-enron-GM5-ALL-ALL-5

| GM5 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | RAN | 0.7371 | 0.9091 | 0.5956 | 0.0726 | 0.4816 | 0.9491 | 0.0000 | 0.3135 |
| | | SAL | 0.7413 | 0.8168 | 0.6452 | 0.0565 | 0.4794 | 0.9396 | 0.0000 | 0.3142 |
| | | STL | 0.5624 | 0.8320 | 0.3708 | 0.1317 | 0.5353 | 0.9870 | 0.0000 | 0.2520 |
| | 1 | RAN | 0.7178 | 0.8903 | 0.5807 | 0.0706 | 0.5411 | 0.9505 | 0.0000 | 0.2318 |
| | | SAL | 0.7020 | 0.8310 | 0.5165 | 0.0841 | 0.5309 | 0.9173 | 0.0000 | 0.2269 |
| | | STL | 0.6472 | 0.7498 | 0.5462 | 0.0643 | 0.6284 | 0.9250 | 0.0000 | 0.1657 |
| | 2 | RAN | 0.7544 | 0.8888 | 0.6724 | 0.0622 | 0.5753 | 0.9466 | 0.0000 | 0.2423 |
| | | SAL | 0.7632 | 0.8734 | 0.6045 | 0.0724 | 0.5702 | 0.9409 | 0.0000 | 0.2430 |
| | | STL | 0.6647 | 0.7891 | 0.5714 | 0.0708 | 0.6460 | 0.9275 | 0.0000 | 0.1618 |
| | 4 | RAN | 0.7575 | 0.8857 | 0.6151 | 0.0698 | 0.5545 | 0.9315 | 0.0000 | 0.2608 |
| | | SAL | 0.7746 | 0.8638 | 0.6282 | 0.0632 | 0.5451 | 0.9359 | 0.0000 | 0.2748 |
| | | STL | 0.6779 | 0.7848 | 0.5526 | 0.0761 | 0.6616 | 0.9444 | 0.0000 | 0.1664 |
| | 8 | RAN | 0.7310 | 0.8473 | 0.5485 | 0.0869 | 0.4717 | 0.9418 | 0.0000 | 0.2702 |
| | | SAL | 0.7329 | 0.8184 | 0.6297 | 0.0669 | 0.4656 | 0.9433 | 0.0000 | 0.2772 |
| | | STL | 0.6869 | 0.8005 | 0.5586 | 0.0830 | 0.6696 | 0.9737 | 0.0000 | 0.1759 |
| | 16 | RAN | 0.7199 | 0.8389 | 0.4791 | 0.0847 | 0.4327 | 0.9548 | 0.0000 | 0.2947 |
| | | SAL | 0.6940 | 0.8021 | 0.5511 | 0.0835 | 0.3992 | 0.9386 | 0.0000 | 0.2790 |
| | | STL | 0.6834 | 0.7931 | 0.5616 | 0.0810 | 0.6685 | 0.9867 | 0.0000 | 0.1765 |

Table G.14: grouped-nb-enron-GM5-ALL-ALL-10

| GM5 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | RAN | 0.6434 | 0.7309 | 0.5130 | 0.0775 | 0.4455 | 0.9297 | 0.0000 | 0.3071 |
| | | SAL | 0.6732 | 0.7168 | 0.5666 | 0.0525 | 0.4418 | 0.9178 | 0.0000 | 0.3076 |
| | | STL | 0.5076 | 0.7289 | 0.4074 | 0.1105 | 0.4916 | 0.9744 | 0.0000 | 0.2698 |
| | 1 | RAN | 0.6028 | 0.7185 | 0.5085 | 0.0748 | 0.4394 | 0.9144 | 0.0000 | 0.2220 |
| | | SAL | 0.5991 | 0.6973 | 0.5529 | 0.0464 | 0.4385 | 0.8812 | 0.0000 | 0.2177 |
| | | STL | 0.5522 | 0.6382 | 0.5079 | 0.0417 | 0.5254 | 0.9268 | 0.0000 | 0.1864 |
| | 2 | RAN | 0.6698 | 0.6963 | 0.6157 | 0.0272 | 0.4917 | 0.8946 | 0.0000 | 0.2253 |
| | | SAL | 0.6673 | 0.7188 | 0.6446 | 0.0239 | 0.4889 | 0.8985 | 0.0000 | 0.2270 |
| | | STL | 0.5702 | 0.6659 | 0.5047 | 0.0519 | 0.5431 | 0.9620 | 0.0000 | 0.1840 |
| | 4 | RAN | 0.6709 | 0.7459 | 0.5991 | 0.0609 | 0.4703 | 0.9324 | 0.0000 | 0.2584 |
| | | SAL | 0.6765 | 0.7014 | 0.6575 | 0.0159 | 0.4689 | 0.8822 | 0.0000 | 0.2598 |
| | | STL | 0.5861 | 0.6627 | 0.4969 | 0.0600 | 0.5599 | 0.9444 | 0.0000 | 0.1940 |
| | 8 | RAN | 0.6294 | 0.7127 | 0.5875 | 0.0421 | 0.3859 | 0.9066 | 0.0000 | 0.2616 |
| | | SAL | 0.5961 | 0.6366 | 0.5724 | 0.0224 | 0.3662 | 0.9130 | 0.0000 | 0.2548 |
| | | STL | 0.5726 | 0.6436 | 0.5000 | 0.0465 | 0.5428 | 0.9867 | 0.0000 | 0.2030 |
| | 16 | RAN | 0.5891 | 0.6076 | 0.5691 | 0.0148 | 0.3203 | 0.9140 | 0.0000 | 0.2634 |
| | | SAL | 0.5466 | 0.5729 | 0.5154 | 0.0211 | 0.2859 | 0.9126 | 0.0000 | 0.2409 |
| | | STL | 0.5575 | 0.6254 | 0.4921 | 0.0406 | 0.5264 | 0.9730 | 0.0000 | 0.2115 |

Table G.15: grouped-nb-enron-GM5-ALL-ALL-25

| GM5 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | RAN | 0.6165 | 0.6977 | 0.5449 | 0.0628 | 0.4160 | 0.9251 | 0.0000 | 0.3096 |
| | | SAL | 0.6224 | 0.6475 | 0.5970 | 0.0206 | 0.4147 | 0.9096 | 0.0000 | 0.3081 |
| | | STL | 0.4838 | 0.6288 | 0.3783 | 0.1060 | 0.4544 | 0.9600 | 0.0000 | 0.2827 |
| | 1 | RAN | 0.5128 | 0.5454 | 0.4620 | 0.0364 | 0.3799 | 0.8700 | 0.0000 | 0.2134 |
| | | SAL | 0.5173 | 0.5303 | 0.4917 | 0.0181 | 0.3783 | 0.8631 | 0.0000 | 0.2125 |
| | | STL | 0.5007 | 0.5410 | 0.4597 | 0.0332 | 0.4600 | 0.9136 | 0.0000 | 0.1971 |
| | 2 | RAN | 0.5942 | 0.6422 | 0.5384 | 0.0427 | 0.4341 | 0.9067 | 0.0000 | 0.2238 |
| | | SAL | 0.5959 | 0.6162 | 0.5765 | 0.0162 | 0.4326 | 0.8889 | 0.0000 | 0.2218 |
| | | STL | 0.5277 | 0.5604 | 0.4907 | 0.0286 | 0.4779 | 0.9041 | 0.0000 | 0.1944 |
| | 4 | RAN | 0.5791 | 0.6349 | 0.4732 | 0.0749 | 0.4192 | 0.8708 | 0.0000 | 0.2509 |
| | | SAL | 0.6179 | 0.6355 | 0.6059 | 0.0127 | 0.4192 | 0.8923 | 0.0000 | 0.2568 |
| | | STL | 0.5356 | 0.5698 | 0.4859 | 0.0360 | 0.4860 | 0.8986 | 0.0000 | 0.2134 |
| | 8 | RAN | 0.5125 | 0.5991 | 0.3952 | 0.0860 | 0.2985 | 0.8772 | 0.0000 | 0.2387 |
| | | SAL | 0.5315 | 0.5655 | 0.5055 | 0.0251 | 0.3084 | 0.9045 | 0.0000 | 0.2436 |
| | | STL | 0.4985 | 0.5234 | 0.4773 | 0.0190 | 0.4363 | 0.8986 | 0.0000 | 0.2272 |
| | 16 | RAN | 0.4494 | 0.5206 | 0.3570 | 0.0685 | 0.2269 | 0.9072 | 0.0000 | 0.2189 |
| | | SAL | 0.4740 | 0.5064 | 0.4419 | 0.0263 | 0.2237 | 0.8982 | 0.0000 | 0.2247 |
| | | STL | 0.4717 | 0.5027 | 0.4357 | 0.0276 | 0.4012 | 0.8986 | 0.0000 | 0.2360 |

Table G.16: grouped-nb-enron-GM5-ALL-ALL-50

| GM5 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | RAN | 0.6083 | 0.6216 | 0.5951 | 0.0132 | 0.4086 | 0.9157 | 0.0000 | 0.3082 |
| | | SAL | 0.5978 | 0.6346 | 0.5609 | 0.0368 | 0.4000 | 0.9017 | 0.0000 | 0.3079 |
| | | STL | 0.4877 | 0.5973 | 0.3780 | 0.1096 | 0.4317 | 0.9444 | 0.0000 | 0.2888 |
| | 1 | RAN | 0.4809 | 0.4897 | 0.4722 | 0.0087 | 0.3491 | 0.8622 | 0.0000 | 0.2108 |
| | | SAL | 0.4856 | 0.5150 | 0.4562 | 0.0294 | 0.3503 | 0.8414 | 0.0000 | 0.2088 |
| | | STL | 0.4500 | 0.4597 | 0.4403 | 0.0097 | 0.4031 | 0.8246 | 0.0000 | 0.1910 |
| | 2 | RAN | 0.5593 | 0.5894 | 0.5291 | 0.0301 | 0.4092 | 0.8707 | 0.0000 | 0.2188 |
| | | SAL | 0.5609 | 0.5746 | 0.5471 | 0.0138 | 0.4055 | 0.8847 | 0.0000 | 0.2155 |
| | | STL | 0.4973 | 0.5170 | 0.4776 | 0.0197 | 0.4377 | 0.8857 | 0.0000 | 0.1953 |
| | 4 | RAN | 0.5787 | 0.5833 | 0.5741 | 0.0046 | 0.3985 | 0.8586 | 0.0000 | 0.2468 |
| | | SAL | 0.5722 | 0.5798 | 0.5647 | 0.0076 | 0.3930 | 0.8831 | 0.0000 | 0.2525 |
| | | STL | 0.5108 | 0.5426 | 0.4791 | 0.0318 | 0.4419 | 0.8720 | 0.0000 | 0.2255 |
| | 8 | RAN | 0.4814 | 0.5279 | 0.4349 | 0.0465 | 0.2771 | 0.9046 | 0.0000 | 0.2352 |
| | | SAL | 0.4896 | 0.5072 | 0.4719 | 0.0176 | 0.2744 | 0.8724 | 0.0000 | 0.2384 |
| | | STL | 0.4664 | 0.4738 | 0.4590 | 0.0074 | 0.3636 | 0.8848 | 0.0000 | 0.2289 |
| | 16 | RAN | 0.4317 | 0.4459 | 0.4176 | 0.0142 | 0.1951 | 0.8947 | 0.0000 | 0.2138 |
| | | SAL | 0.4290 | 0.4481 | 0.4100 | 0.0191 | 0.1969 | 0.8697 | 0.0000 | 0.2133 |
| | | STL | 0.4339 | 0.4690 | 0.3988 | 0.0351 | 0.3157 | 0.8950 | 0.0000 | 0.2352 |

Table G.17: grouped-nb-enron-GM5-ALL-ALL-75

| GM5 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | RAN | 0.5661 | 0.5661 | 0.5661 | 0.0000 | 0.3832 | 0.9077 | 0.0000 | 0.3073 |
| | | SAL | 0.5657 | 0.5657 | 0.5657 | 0.0000 | 0.3823 | 0.9085 | 0.0000 | 0.3065 |
| | | STL | 0.5657 | 0.5657 | 0.5657 | 0.0000 | 0.3823 | 0.9085 | 0.0000 | 0.3065 |
| | 1 | RAN | 0.4141 | 0.4141 | 0.4141 | 0.0000 | 0.3101 | 0.8081 | 0.0000 | 0.2001 |
| | | SAL | 0.4137 | 0.4137 | 0.4137 | 0.0000 | 0.3097 | 0.8095 | 0.0000 | 0.1995 |
| | | STL | 0.4137 | 0.4137 | 0.4137 | 0.0000 | 0.3097 | 0.8095 | 0.0000 | 0.1995 |
| | 2 | RAN | 0.4945 | 0.4945 | 0.4945 | 0.0000 | 0.3653 | 0.8565 | 0.0000 | 0.2067 |
| | | SAL | 0.4938 | 0.4938 | 0.4938 | 0.0000 | 0.3645 | 0.8586 | 0.0000 | 0.2073 |
| | | STL | 0.4938 | 0.4938 | 0.4938 | 0.0000 | 0.3645 | 0.8586 | 0.0000 | 0.2073 |
| | 4 | RAN | 0.5127 | 0.5127 | 0.5127 | 0.0000 | 0.3546 | 0.8444 | 0.0000 | 0.2496 |
| | | SAL | 0.5125 | 0.5125 | 0.5125 | 0.0000 | 0.3545 | 0.8440 | 0.0000 | 0.2489 |
| | | STL | 0.5125 | 0.5125 | 0.5125 | 0.0000 | 0.3545 | 0.8440 | 0.0000 | 0.2489 |
| | 8 | RAN | 0.4285 | 0.4285 | 0.4285 | 0.0000 | 0.2320 | 0.8633 | 0.0000 | 0.2195 |
| | | SAL | 0.4287 | 0.4287 | 0.4287 | 0.0000 | 0.2322 | 0.8650 | 0.0000 | 0.2195 |
| | | STL | 0.4287 | 0.4287 | 0.4287 | 0.0000 | 0.2322 | 0.8650 | 0.0000 | 0.2195 |
| | 16 | RAN | 0.0433 | 0.0433 | 0.0433 | 0.0000 | 0.0024 | 0.1148 | 0.0000 | 0.0130 |
| | | SAL | 0.3703 | 0.3703 | 0.3703 | 0.0000 | 0.1583 | 0.8920 | 0.0000 | 0.1919 |
| | | STL | 0.3703 | 0.3703 | 0.3703 | 0.0000 | 0.1583 | 0.8920 | 0.0000 | 0.1919 |

Table G.18: grouped-nb-enron-GM5-ALL-ALL-150

| GB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | RAN | 0.7853 | 0.9709 | 0.5772 | 0.0870 | 0.5006 | 0.9852 | 0.0000 | 0.3368 |
| | | SAL | 0.7730 | 0.9016 | 0.6315 | 0.0687 | 0.4223 | 0.9392 | 0.0000 | 0.3270 |
| | | STL | 0.8062 | 0.9319 | 0.6075 | 0.0863 | 0.7812 | 0.9660 | 0.0000 | 0.1475 |
| | 1 | RAN | 0.8290 | 0.9432 | 0.5311 | 0.1033 | 0.6821 | 0.9757 | 0.0000 | 0.2415 |
| | | SAL | 0.8449 | 0.9561 | 0.4360 | 0.1173 | 0.6614 | 0.9776 | 0.0000 | 0.2596 |
| | | STL | 0.7761 | 0.9485 | 0.4815 | 0.1023 | 0.7527 | 0.9776 | 0.0000 | 0.1564 |
| | 2 | RAN | 0.8492 | 0.9669 | 0.5602 | 0.1006 | 0.6945 | 0.9833 | 0.0000 | 0.2540 |
| | | SAL | 0.8578 | 0.9550 | 0.4790 | 0.1102 | 0.6722 | 0.9820 | 0.0000 | 0.2637 |
| | | STL | 0.7874 | 0.9494 | 0.5556 | 0.0891 | 0.7623 | 0.9757 | 0.0000 | 0.1571 |
| | 4 | RAN | 0.8901 | 0.9483 | 0.8122 | 0.0359 | 0.7044 | 0.9757 | 0.0000 | 0.2564 |
| | | SAL | 0.8930 | 0.9522 | 0.7851 | 0.0455 | 0.6939 | 0.9823 | 0.0000 | 0.2622 |
| | | STL | 0.8055 | 0.9459 | 0.5556 | 0.0870 | 0.7835 | 0.9763 | 0.0000 | 0.1410 |
| | 8 | RAN | 0.8854 | 0.9782 | 0.8101 | 0.0419 | 0.6996 | 0.9889 | 0.0000 | 0.2607 |
| | | SAL | 0.8904 | 0.9579 | 0.7796 | 0.0422 | 0.6893 | 0.9870 | 0.0000 | 0.2652 |
| | | STL | 0.8045 | 0.9528 | 0.5556 | 0.0849 | 0.7806 | 0.9771 | 0.0000 | 0.1456 |
| | 16 | RAN | 0.8777 | 0.9674 | 0.7354 | 0.0531 | 0.6874 | 1.0000 | 0.0000 | 0.2686 |
| | | SAL | 0.8915 | 0.9528 | 0.7770 | 0.0457 | 0.6786 | 0.9870 | 0.0000 | 0.2758 |
| | | STL | 0.8075 | 0.9473 | 0.5556 | 0.0903 | 0.7863 | 0.9765 | 0.0000 | 0.1467 |

Table G.19: grouped-nb-enron-GB3-ALL-ALL-5

| GB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | RAN | 0.7209 | 0.8310 | 0.5989 | 0.0743 | 0.3650 | 0.9456 | 0.0000 | 0.3224 |
| | | SAL | 0.6704 | 0.8012 | 0.5652 | 0.0738 | 0.3254 | 0.9493 | 0.0000 | 0.2913 |
| | | STL | 0.7258 | 0.8951 | 0.5319 | 0.0953 | 0.6902 | 0.9610 | 0.0000 | 0.2041 |
| | 1 | RAN | 0.7835 | 0.8924 | 0.5487 | 0.1081 | 0.6067 | 0.9558 | 0.0000 | 0.2524 |
| | | SAL | 0.7793 | 0.8962 | 0.5359 | 0.1160 | 0.5999 | 0.9579 | 0.0000 | 0.2547 |
| | | STL | 0.7129 | 0.8734 | 0.5638 | 0.0755 | 0.6784 | 0.9519 | 0.0000 | 0.1842 |
| | 2 | RAN | 0.7969 | 0.9087 | 0.5466 | 0.1090 | 0.6218 | 0.9724 | 0.0000 | 0.2542 |
| | | SAL | 0.7994 | 0.8989 | 0.5161 | 0.1149 | 0.6141 | 0.9637 | 0.0000 | 0.2592 |
| | | STL | 0.7225 | 0.9138 | 0.5532 | 0.0917 | 0.6887 | 0.9594 | 0.0000 | 0.1963 |
| | 4 | RAN | 0.8314 | 0.9251 | 0.7516 | 0.0528 | 0.6422 | 0.9688 | 0.0000 | 0.2545 |
| | | SAL | 0.8502 | 0.9004 | 0.7678 | 0.0372 | 0.6340 | 0.9648 | 0.0000 | 0.2669 |
| | | STL | 0.7395 | 0.9189 | 0.5532 | 0.0938 | 0.7087 | 0.9663 | 0.0000 | 0.1855 |
| | 8 | RAN | 0.8343 | 0.9456 | 0.7188 | 0.0573 | 0.6320 | 1.0000 | 0.0000 | 0.2648 |
| | | SAL | 0.8504 | 0.9029 | 0.7455 | 0.0378 | 0.6338 | 0.9744 | 0.0000 | 0.2681 |
| | | STL | 0.7374 | 0.9109 | 0.5638 | 0.0883 | 0.7052 | 0.9772 | 0.0000 | 0.1874 |
| | 16 | RAN | 0.8336 | 0.9165 | 0.6203 | 0.0721 | 0.6227 | 0.9632 | 0.0000 | 0.2742 |
| | | SAL | 0.8500 | 0.8999 | 0.7729 | 0.0369 | 0.6262 | 0.9744 | 0.0000 | 0.2755 |
| | | STL | 0.7437 | 0.9198 | 0.5426 | 0.0988 | 0.7127 | 0.9686 | 0.0000 | 0.1909 |

Table G.20: grouped-nb-enron-GB3-ALL-ALL-10

| GB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | RAN | 0.6238 | 0.7052 | 0.5110 | 0.0638 | 0.2828 | 0.9304 | 0.0000 | 0.2843 |
| | | SAL | 0.5608 | 0.6636 | 0.4270 | 0.0765 | 0.2493 | 0.9253 | 0.0000 | 0.2652 |
| | | STL | 0.6151 | 0.6894 | 0.5403 | 0.0542 | 0.5556 | 0.9600 | 0.0000 | 0.2281 |
| | 1 | RAN | 0.7123 | 0.8258 | 0.6218 | 0.0757 | 0.5320 | 0.9329 | 0.0000 | 0.2465 |
| | | SAL | 0.7184 | 0.8183 | 0.5719 | 0.0860 | 0.5347 | 0.9561 | 0.0000 | 0.2462 |
| | | STL | 0.6352 | 0.6956 | 0.5542 | 0.0558 | 0.5955 | 0.9383 | 0.0000 | 0.2017 |
| | 2 | RAN | 0.7365 | 0.8493 | 0.6077 | 0.0752 | 0.5446 | 0.9415 | 0.0000 | 0.2548 |
| | | SAL | 0.7453 | 0.8321 | 0.6359 | 0.0740 | 0.5483 | 0.9409 | 0.0000 | 0.2533 |
| | | STL | 0.6487 | 0.7517 | 0.5610 | 0.0697 | 0.6071 | 0.9620 | 0.0000 | 0.2077 |
| | 4 | RAN | 0.7956 | 0.8607 | 0.7353 | 0.0444 | 0.5743 | 0.9528 | 0.0000 | 0.2585 |
| | | SAL | 0.7939 | 0.8307 | 0.7504 | 0.0261 | 0.5702 | 0.9514 | 0.0000 | 0.2595 |
| | | STL | 0.6663 | 0.8136 | 0.5501 | 0.0905 | 0.6264 | 0.9620 | 0.0000 | 0.2075 |
| | 8 | RAN | 0.7805 | 0.8614 | 0.7359 | 0.0430 | 0.5716 | 0.9575 | 0.0000 | 0.2592 |
| | | SAL | 0.7891 | 0.8112 | 0.7535 | 0.0199 | 0.5687 | 0.9647 | 0.0000 | 0.2588 |
| | | STL | 0.6612 | 0.8189 | 0.5528 | 0.0916 | 0.6221 | 0.9690 | 0.0000 | 0.2107 |
| | 16 | RAN | 0.7973 | 0.8677 | 0.7348 | 0.0417 | 0.5712 | 0.9620 | 0.0000 | 0.2693 |
| | | SAL | 0.8016 | 0.8327 | 0.7752 | 0.0195 | 0.5698 | 0.9465 | 0.0000 | 0.2712 |
| | | STL | 0.6681 | 0.8462 | 0.5556 | 0.1027 | 0.6307 | 0.9500 | 0.0000 | 0.2141 |

Table G.21: grouped-nb-enron-GB3-ALL-ALL-25

| GB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | RAN | 0.5052 | 0.5421 | 0.4843 | 0.0262 | 0.1996 | 0.8947 | 0.0000 | 0.2422 |
| | | SAL | 0.5059 | 0.5463 | 0.4256 | 0.0567 | 0.2082 | 0.8943 | 0.0000 | 0.2474 |
| | | STL | 0.5067 | 0.5326 | 0.4752 | 0.0238 | 0.4126 | 0.9444 | 0.0000 | 0.2554 |
| | 1 | RAN | 0.6608 | 0.7451 | 0.5709 | 0.0712 | 0.4806 | 0.9394 | 0.0000 | 0.2483 |
| | | SAL | 0.6613 | 0.7140 | 0.6267 | 0.0379 | 0.4868 | 0.9222 | 0.0000 | 0.2433 |
| | | STL | 0.5952 | 0.6454 | 0.5289 | 0.0489 | 0.5351 | 0.9077 | 0.0000 | 0.2158 |
| | 2 | RAN | 0.7032 | 0.7629 | 0.6679 | 0.0425 | 0.5026 | 0.9152 | 0.0000 | 0.2543 |
| | | SAL | 0.7040 | 0.7769 | 0.6296 | 0.0602 | 0.5069 | 0.9456 | 0.0000 | 0.2517 |
| | | STL | 0.6148 | 0.7059 | 0.5262 | 0.0734 | 0.5487 | 0.9620 | 0.0000 | 0.2229 |
| | 4 | RAN | 0.7465 | 0.8163 | 0.6789 | 0.0561 | 0.5280 | 0.9539 | 0.0000 | 0.2619 |
| | | SAL | 0.7485 | 0.7770 | 0.7042 | 0.0317 | 0.5268 | 0.9419 | 0.0000 | 0.2586 |
| | | STL | 0.6341 | 0.7525 | 0.5340 | 0.0902 | 0.5674 | 0.9467 | 0.0000 | 0.2210 |
| | 8 | RAN | 0.7456 | 0.7993 | 0.6834 | 0.0477 | 0.5262 | 0.9637 | 0.0000 | 0.2594 |
| | | SAL | 0.7495 | 0.7739 | 0.7146 | 0.0253 | 0.5304 | 0.9650 | 0.0000 | 0.2577 |
| | | STL | 0.6298 | 0.7552 | 0.5255 | 0.0950 | 0.5652 | 0.9637 | 0.0000 | 0.2239 |
| | 16 | RAN | 0.7533 | 0.8277 | 0.7077 | 0.0531 | 0.5326 | 0.9410 | 0.0000 | 0.2718 |
| | | SAL | 0.7629 | 0.7801 | 0.7349 | 0.0200 | 0.5322 | 0.9392 | 0.0000 | 0.2685 |
| | | STL | 0.6392 | 0.7742 | 0.5201 | 0.1044 | 0.5761 | 0.9744 | 0.0000 | 0.2270 |

Table G.22: grouped-nb-enron-GB3-ALL-ALL-50

| GB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | RAN | 0.4676 | 0.4949 | 0.4403 | 0.0273 | 0.1844 | 0.8811 | 0.0000 | 0.2307 |
| | | SAL | 0.4682 | 0.4766 | 0.4598 | 0.0084 | 0.1812 | 0.8809 | 0.0000 | 0.2314 |
| | | STL | 0.4805 | 0.4962 | 0.4647 | 0.0157 | 0.3218 | 0.8709 | 0.0000 | 0.2559 |
| | 1 | RAN | 0.6370 | 0.7176 | 0.5564 | 0.0806 | 0.4643 | 0.9202 | 0.0000 | 0.2460 |
| | | SAL | 0.6396 | 0.6479 | 0.6313 | 0.0083 | 0.4636 | 0.9209 | 0.0000 | 0.2436 |
| | | STL | 0.5797 | 0.6272 | 0.5323 | 0.0474 | 0.4966 | 0.9021 | 0.0000 | 0.2181 |
| | 2 | RAN | 0.6822 | 0.7352 | 0.6293 | 0.0529 | 0.4818 | 0.9308 | 0.0000 | 0.2558 |
| | | SAL | 0.6820 | 0.7254 | 0.6387 | 0.0433 | 0.4829 | 0.9281 | 0.0000 | 0.2524 |
| | | STL | 0.6076 | 0.6827 | 0.5324 | 0.0752 | 0.5153 | 0.9301 | 0.0000 | 0.2252 |
| | 4 | RAN | 0.7239 | 0.7331 | 0.7147 | 0.0092 | 0.5013 | 0.9378 | 0.0000 | 0.2647 |
| | | SAL | 0.7261 | 0.7345 | 0.7178 | 0.0083 | 0.5004 | 0.9382 | 0.0000 | 0.2623 |
| | | STL | 0.6296 | 0.7232 | 0.5359 | 0.0936 | 0.5338 | 0.9385 | 0.0000 | 0.2288 |
| | 8 | RAN | 0.7270 | 0.7491 | 0.7049 | 0.0221 | 0.5057 | 0.9615 | 0.0000 | 0.2607 |
| | | SAL | 0.7284 | 0.7350 | 0.7217 | 0.0067 | 0.5036 | 0.9550 | 0.0000 | 0.2627 |
| | | STL | 0.6302 | 0.7245 | 0.5359 | 0.0943 | 0.5350 | 0.9597 | 0.0000 | 0.2300 |
| | 16 | RAN | 0.7417 | 0.7559 | 0.7275 | 0.0142 | 0.5076 | 0.9369 | 0.0000 | 0.2714 |
| | | SAL | 0.7437 | 0.7479 | 0.7395 | 0.0042 | 0.5091 | 0.9250 | 0.0000 | 0.2735 |
| | | STL | 0.6370 | 0.7428 | 0.5311 | 0.1059 | 0.5410 | 0.9500 | 0.0000 | 0.2348 |

Table G.23: grouped-nb-enron-GB3-ALL-ALL-75

| GB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | RAN | 0.4279 | 0.4279 | 0.4279 | 0.0000 | 0.1609 | 0.8518 | 0.0000 | 0.2174 |
| | | SAL | 0.4284 | 0.4284 | 0.4284 | 0.0000 | 0.1612 | 0.8550 | 0.0000 | 0.2176 |
| | | STL | 0.4284 | 0.4284 | 0.4284 | 0.0000 | 0.1612 | 0.8550 | 0.0000 | 0.2176 |
| | 1 | RAN | 0.5971 | 0.5971 | 0.5971 | 0.0000 | 0.4266 | 0.8973 | 0.0000 | 0.2455 |
| | | SAL | 0.5978 | 0.5978 | 0.5978 | 0.0000 | 0.4264 | 0.8962 | 0.0000 | 0.2449 |
| | | STL | 0.5978 | 0.5978 | 0.5978 | 0.0000 | 0.4264 | 0.8962 | 0.0000 | 0.2449 |
| | 2 | RAN | 0.6499 | 0.6499 | 0.6499 | 0.0000 | 0.4491 | 0.9160 | 0.0000 | 0.2541 |
| | | SAL | 0.6499 | 0.6499 | 0.6499 | 0.0000 | 0.4498 | 0.9190 | 0.0000 | 0.2528 |
| | | STL | 0.6499 | 0.6499 | 0.6499 | 0.0000 | 0.4498 | 0.9190 | 0.0000 | 0.2528 |
| | 4 | RAN | 0.6862 | 0.6862 | 0.6862 | 0.0000 | 0.4657 | 0.9348 | 0.0000 | 0.2595 |
| | | SAL | 0.6859 | 0.6859 | 0.6859 | 0.0000 | 0.4644 | 0.9331 | 0.0000 | 0.2612 |
| | | STL | 0.6859 | 0.6859 | 0.6859 | 0.0000 | 0.4644 | 0.9331 | 0.0000 | 0.2612 |
| | 8 | RAN | 0.6885 | 0.6885 | 0.6885 | 0.0000 | 0.4703 | 0.9593 | 0.0000 | 0.2593 |
| | | SAL | 0.6891 | 0.6891 | 0.6891 | 0.0000 | 0.4700 | 0.9583 | 0.0000 | 0.2604 |
| | | STL | 0.6891 | 0.6891 | 0.6891 | 0.0000 | 0.4700 | 0.9583 | 0.0000 | 0.2604 |
| | 16 | RAN | 0.7052 | 0.7052 | 0.7052 | 0.0000 | 0.4745 | 0.9287 | 0.0000 | 0.2686 |
| | | SAL | 0.7059 | 0.7059 | 0.7059 | 0.0000 | 0.4752 | 0.9278 | 0.0000 | 0.2689 |
| | | STL | 0.7059 | 0.7059 | 0.7059 | 0.0000 | 0.4752 | 0.9278 | 0.0000 | 0.2689 |

Table G.24: grouped-nb-enron-GB3-ALL-ALL-150

| OSB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | Group Type | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | RAN | 0.8658 | 0.9592 | 0.7085 | 0.0647 | 0.6505 | 1.0000 | 0.0000 | 0.2854 |
| | | SAL | 0.8678 | 0.9558 | 0.7158 | 0.0494 | 0.6314 | 1.0000 | 0.0000 | 0.2889 |
| | | STL | 0.8244 | 0.9564 | 0.5185 | 0.0956 | 0.8052 | 0.9798 | 0.0000 | 0.1462 |
| | 1 | RAN | 0.8825 | 0.9533 | 0.7244 | 0.0602 | 0.7007 | 1.0000 | 0.0000 | 0.2649 |
| | | SAL | 0.8912 | 0.9566 | 0.7746 | 0.0434 | 0.6783 | 0.9832 | 0.0000 | 0.2763 |
| | | STL | 0.8207 | 0.9574 | 0.5556 | 0.0896 | 0.8008 | 0.9781 | 0.0000 | 0.1418 |
| | 2 | RAN | 0.8786 | 0.9398 | 0.7018 | 0.0585 | 0.7039 | 0.9870 | 0.0000 | 0.2643 |
| | | SAL | 0.8928 | 0.9575 | 0.7781 | 0.0397 | 0.6756 | 0.9835 | 0.0000 | 0.2781 |
| | | STL | 0.8210 | 0.9487 | 0.5556 | 0.0915 | 0.8013 | 0.9785 | 0.0000 | 0.1429 |
| | 4 | RAN | 0.8892 | 0.9587 | 0.7664 | 0.0446 | 0.6944 | 0.9773 | 0.0000 | 0.2691 |
| | | SAL | 0.8930 | 0.9578 | 0.7772 | 0.0400 | 0.6759 | 0.9870 | 0.0000 | 0.2784 |
| | | STL | 0.8211 | 0.9560 | 0.5556 | 0.0914 | 0.8012 | 0.9789 | 0.0000 | 0.1430 |
| | 8 | RAN | 0.8764 | 0.9714 | 0.7461 | 0.0546 | 0.6986 | 0.9881 | 0.0000 | 0.2581 |
| | | SAL | 0.8932 | 0.9582 | 0.7772 | 0.0411 | 0.6767 | 1.0000 | 0.0000 | 0.2786 |
| | | STL | 0.8219 | 0.9564 | 0.5556 | 0.0920 | 0.8026 | 0.9786 | 0.0000 | 0.1431 |
| | 16 | RAN | 0.8806 | 0.9823 | 0.7683 | 0.0558 | 0.7004 | 0.9913 | 0.0000 | 0.2600 |
| | | SAL | 0.8930 | 0.9582 | 0.7763 | 0.0401 | 0.6752 | 1.0000 | 0.0000 | 0.2794 |
| | | STL | 0.8224 | 0.9564 | 0.5556 | 0.0922 | 0.8028 | 0.9786 | 0.0000 | 0.1433 |

Table G.25: grouped-nb-enron-OSB3-ALL-ALL-5

| OSB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | Group Type | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | RAN | 0.8152 | 0.9272 | 0.6736 | 0.0637 | 0.5782 | 0.9867 | 0.0000 | 0.2980 |
| | | SAL | 0.8133 | 0.8925 | 0.6713 | 0.0544 | 0.5763 | 1.0000 | 0.0000 | 0.2895 |
| | | STL | 0.7616 | 0.9336 | 0.5532 | 0.1020 | 0.7350 | 0.9722 | 0.0000 | 0.1888 |
| | 1 | RAN | 0.8423 | 0.9261 | 0.7550 | 0.0474 | 0.6266 | 0.9744 | 0.0000 | 0.2812 |
| | | SAL | 0.8511 | 0.9143 | 0.7614 | 0.0386 | 0.6205 | 0.9635 | 0.0000 | 0.2861 |
| | | STL | 0.7549 | 0.9317 | 0.5213 | 0.1034 | 0.7281 | 0.9682 | 0.0000 | 0.1880 |
| | 2 | RAN | 0.8472 | 0.9178 | 0.7590 | 0.0516 | 0.6301 | 0.9731 | 0.0000 | 0.2804 |
| | | SAL | 0.8514 | 0.9140 | 0.7625 | 0.0392 | 0.6190 | 0.9636 | 0.0000 | 0.2867 |
| | | STL | 0.7555 | 0.9272 | 0.5213 | 0.1063 | 0.7297 | 0.9667 | 0.0000 | 0.1886 |
| | 4 | RAN | 0.8447 | 0.9260 | 0.7648 | 0.0470 | 0.6368 | 0.9655 | 0.0000 | 0.2737 |
| | | SAL | 0.8526 | 0.9149 | 0.7624 | 0.0389 | 0.6195 | 0.9637 | 0.0000 | 0.2878 |
| | | STL | 0.7557 | 0.9307 | 0.5213 | 0.1060 | 0.7291 | 0.9685 | 0.0000 | 0.1895 |
| | 8 | RAN | 0.8492 | 0.9332 | 0.7598 | 0.0492 | 0.6363 | 0.9744 | 0.0000 | 0.2794 |
| | | SAL | 0.8524 | 0.9152 | 0.7617 | 0.0382 | 0.6194 | 0.9744 | 0.0000 | 0.2876 |
| | | STL | 0.7570 | 0.9315 | 0.5213 | 0.1071 | 0.7312 | 0.9694 | 0.0000 | 0.1889 |
| | 16 | RAN | 0.8458 | 0.9123 | 0.7765 | 0.0452 | 0.6265 | 0.9870 | 0.0000 | 0.2801 |
| | | SAL | 0.8525 | 0.9154 | 0.7615 | 0.0388 | 0.6191 | 0.9870 | 0.0000 | 0.2881 |
| | | STL | 0.7574 | 0.9318 | 0.5213 | 0.1073 | 0.7315 | 0.9694 | 0.0000 | 0.1891 |

Table G.26: grouped-nb-enron-OSB3-ALL-ALL-10

| OSB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | RAN | 0.7452 | 0.8205 | 0.6445 | 0.0623 | 0.5169 | 1.0000 | 0.0000 | 0.2949 |
| | | SAL | 0.7481 | 0.8048 | 0.6397 | 0.0565 | 0.5206 | 0.9867 | 0.0000 | 0.2891 |
| | | STL | 0.6854 | 0.8437 | 0.5711 | 0.1006 | 0.6488 | 0.9794 | 0.0000 | 0.2163 |
| | 1 | RAN | 0.7937 | 0.8378 | 0.7226 | 0.0378 | 0.5639 | 0.9870 | 0.0000 | 0.2906 |
| | | SAL | 0.7999 | 0.8301 | 0.7640 | 0.0205 | 0.5692 | 0.9349 | 0.0000 | 0.2831 |
| | | STL | 0.6822 | 0.8489 | 0.5514 | 0.1050 | 0.6463 | 0.9744 | 0.0000 | 0.2137 |
| | 2 | RAN | 0.8030 | 0.8257 | 0.7623 | 0.0212 | 0.5716 | 0.9445 | 0.0000 | 0.2858 |
| | | SAL | 0.8027 | 0.8350 | 0.7624 | 0.0243 | 0.5691 | 0.9386 | 0.0000 | 0.2848 |
| | | STL | 0.6846 | 0.8600 | 0.5514 | 0.1078 | 0.6483 | 0.9744 | 0.0000 | 0.2146 |
| | 4 | RAN | 0.8015 | 0.8546 | 0.7614 | 0.0285 | 0.5681 | 0.9620 | 0.0000 | 0.2803 |
| | | SAL | 0.8010 | 0.8349 | 0.7537 | 0.0268 | 0.5689 | 0.9326 | 0.0000 | 0.2844 |
| | | STL | 0.6839 | 0.8551 | 0.5514 | 0.1067 | 0.6471 | 0.9744 | 0.0000 | 0.2149 |
| | 8 | RAN | 0.7950 | 0.8534 | 0.7415 | 0.0337 | 0.5675 | 0.9744 | 0.0000 | 0.2823 |
| | | SAL | 0.8028 | 0.8354 | 0.7625 | 0.0243 | 0.5681 | 0.9366 | 0.0000 | 0.2867 |
| | | STL | 0.6860 | 0.8618 | 0.5556 | 0.1077 | 0.6498 | 0.9744 | 0.0000 | 0.2140 |
| | 16 | RAN | 0.7977 | 0.8377 | 0.7485 | 0.0324 | 0.5689 | 0.9744 | 0.0000 | 0.2881 |
| | | SAL | 0.8034 | 0.8359 | 0.7622 | 0.0248 | 0.5686 | 0.9384 | 0.0000 | 0.2867 |
| | | STL | 0.6867 | 0.8627 | 0.5556 | 0.1083 | 0.6504 | 0.9744 | 0.0000 | 0.2144 |

Table G.27: grouped-nb-enron-OSB3-ALL-ALL-25

| OSB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | RAN | 0.7132 | 0.7408 | 0.6774 | 0.0265 | 0.4868 | 0.9730 | 0.0000 | 0.2978 |
| | | SAL | 0.7089 | 0.7570 | 0.6429 | 0.0483 | 0.4869 | 0.9730 | 0.0000 | 0.2922 |
| | | STL | 0.6234 | 0.7305 | 0.5190 | 0.0864 | 0.5579 | 1.0000 | 0.0000 | 0.2457 |
| | 1 | RAN | 0.4410 | 0.4869 | 0.3615 | 0.0565 | 0.3216 | 0.8700 | 0.0000 | 0.2438 |
| | | SAL | 0.4123 | 0.4347 | 0.3781 | 0.0246 | 0.3063 | 0.8458 | 0.0000 | 0.2268 |
| | | STL | 0.6105 | 0.6819 | 0.5204 | 0.0672 | 0.5550 | 0.9744 | 0.0000 | 0.2296 |
| | 2 | RAN | 0.4316 | 0.5087 | 0.3695 | 0.0578 | 0.3103 | 0.8427 | 0.0000 | 0.2371 |
| | | SAL | 0.4273 | 0.4539 | 0.3784 | 0.0346 | 0.3078 | 0.8506 | 0.0000 | 0.2314 |
| | | STL | 0.3779 | 0.4000 | 0.3526 | 0.0195 | 0.3548 | 0.9730 | 0.0000 | 0.2091 |
| | 4 | RAN | 0.7562 | 0.7820 | 0.7072 | 0.0346 | 0.5324 | 0.9320 | 0.0000 | 0.2844 |
| | | SAL | 0.7651 | 0.7899 | 0.7354 | 0.0225 | 0.5311 | 0.9343 | 0.0000 | 0.2854 |
| | | STL | 0.6437 | 0.7829 | 0.5197 | 0.1080 | 0.5814 | 0.9744 | 0.0000 | 0.2344 |

Table G.28: grouped-nb-enron-OSB3-ALL-ALL-50

| OSB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | RAN | 0.6983 | 0.6992 | 0.6975 | 0.0009 | 0.4726 | 1.0000 | 0.0000 | 0.3006 |
| | | SAL | 0.6939 | 0.7004 | 0.6874 | 0.0065 | 0.4724 | 0.9867 | 0.0000 | 0.2986 |
| | | STL | 0.6216 | 0.6927 | 0.5504 | 0.0712 | 0.5160 | 0.9870 | 0.0000 | 0.2620 |
| | 1 | RAN | 0.7370 | 0.7448 | 0.7293 | 0.0077 | 0.5072 | 0.9287 | 0.0000 | 0.2889 |
| | | SAL | 0.7425 | 0.7444 | 0.7407 | 0.0018 | 0.5108 | 0.9094 | 0.0000 | 0.2868 |
| | | STL | 0.6427 | 0.7478 | 0.5376 | 0.1051 | 0.5481 | 0.9181 | 0.0000 | 0.2445 |
| | 2 | RAN | 0.7408 | 0.7414 | 0.7402 | 0.0006 | 0.5051 | 0.9341 | 0.0000 | 0.2895 |
| | | SAL | 0.7467 | 0.7488 | 0.7447 | 0.0020 | 0.5117 | 0.9101 | 0.0000 | 0.2873 |
| | | STL | 0.6461 | 0.7531 | 0.5390 | 0.1070 | 0.5486 | 0.9197 | 0.0000 | 0.2461 |
| | 4 | RAN | 0.7378 | 0.7621 | 0.7135 | 0.0243 | 0.5014 | 0.9352 | 0.0000 | 0.2906 |
| | | SAL | 0.7444 | 0.7475 | 0.7412 | 0.0032 | 0.5112 | 0.9077 | 0.0000 | 0.2862 |
| | | STL | 0.6446 | 0.7500 | 0.5393 | 0.1054 | 0.5471 | 0.9208 | 0.0000 | 0.2460 |
| | 8 | RAN | 0.7295 | 0.7860 | 0.6730 | 0.0565 | 0.5098 | 0.9290 | 0.0000 | 0.2863 |
| | | SAL | 0.7477 | 0.7512 | 0.7441 | 0.0036 | 0.5109 | 0.9119 | 0.0000 | 0.2886 |
| | | STL | 0.6465 | 0.7541 | 0.5389 | 0.1076 | 0.5484 | 0.9193 | 0.0000 | 0.2463 |
| | 16 | RAN | 0.7489 | 0.7604 | 0.7374 | 0.0115 | 0.5121 | 0.9315 | 0.0000 | 0.2889 |
| | | SAL | 0.7485 | 0.7515 | 0.7455 | 0.0030 | 0.5117 | 0.9131 | 0.0000 | 0.2887 |
| | | STL | 0.6471 | 0.7551 | 0.5391 | 0.1080 | 0.5491 | 0.9223 | 0.0000 | 0.2467 |

Table G.29: grouped-nb-enron-OSB3-ALL-ALL-75

| OSB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | RAN | 0.6574 | 0.6574 | 0.6574 | 0.0000 | 0.4442 | 0.9867 | 0.0000 | 0.3023 |
| | | SAL | 0.6571 | 0.6571 | 0.6571 | 0.0000 | 0.4441 | 0.9730 | 0.0000 | 0.3023 |
| | | STL | 0.6571 | 0.6571 | 0.6571 | 0.0000 | 0.4441 | 0.9730 | 0.0000 | 0.3023 |
| | 1 | RAN | 0.7040 | 0.7040 | 0.7040 | 0.0000 | 0.4788 | 0.9127 | 0.0000 | 0.2863 |
| | | SAL | 0.7042 | 0.7042 | 0.7042 | 0.0000 | 0.4790 | 0.9118 | 0.0000 | 0.2851 |
| | | STL | 0.7042 | 0.7042 | 0.7042 | 0.0000 | 0.4790 | 0.9118 | 0.0000 | 0.2851 |
| | 2 | RAN | 0.7089 | 0.7089 | 0.7089 | 0.0000 | 0.4786 | 0.9126 | 0.0000 | 0.2879 |
| | | SAL | 0.7092 | 0.7092 | 0.7092 | 0.0000 | 0.4787 | 0.9146 | 0.0000 | 0.2873 |
| | | STL | 0.7092 | 0.7092 | 0.7092 | 0.0000 | 0.4787 | 0.9146 | 0.0000 | 0.2873 |
| | 4 | RAN | 0.7071 | 0.7071 | 0.7071 | 0.0000 | 0.4767 | 0.9189 | 0.0000 | 0.2885 |
| | | SAL | 0.7066 | 0.7066 | 0.7066 | 0.0000 | 0.4770 | 0.9154 | 0.0000 | 0.2869 |
| | | STL | 0.7066 | 0.7066 | 0.7066 | 0.0000 | 0.4770 | 0.9154 | 0.0000 | 0.2869 |
| | 8 | RAN | 0.7100 | 0.7100 | 0.7100 | 0.0000 | 0.4769 | 0.9130 | 0.0000 | 0.2881 |
| | | SAL | 0.7101 | 0.7101 | 0.7101 | 0.0000 | 0.4782 | 0.9151 | 0.0000 | 0.2877 |
| | | STL | 0.7101 | 0.7101 | 0.7101 | 0.0000 | 0.4782 | 0.9151 | 0.0000 | 0.2877 |
| | 16 | RAN | 0.7120 | 0.7120 | 0.7120 | 0.0000 | 0.4793 | 0.9187 | 0.0000 | 0.2893 |
| | | SAL | 0.7112 | 0.7112 | 0.7112 | 0.0000 | 0.4789 | 0.9175 | 0.0000 | 0.2881 |
| | | STL | 0.7112 | 0.7112 | 0.7112 | 0.0000 | 0.4789 | 0.9175 | 0.0000 | 0.2881 |

Table G.30: grouped-nb-enron-OSB3-ALL-ALL-150

# APPENDIX H:
# Grouped Naive Bayes Results for the Twitter Short Message Corpus

| GM1 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | Group Type | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | RAN | 0.6173 | 0.7548 | 0.4740 | 0.0779 | 0.5706 | 0.9069 | 0.0000 | 0.1942 |
| | | SAL | 0.5952 | 0.7598 | 0.4580 | 0.0747 | 0.5405 | 0.9063 | 0.0000 | 0.2002 |
| | | STL | 0.6667 | 0.7714 | 0.5326 | 0.0678 | 0.6548 | 0.9618 | 0.2333 | 0.1351 |
| | 1 | RAN | 0.5577 | 0.7500 | 0.4421 | 0.0738 | 0.5255 | 0.9178 | 0.1455 | 0.1516 |
| | | SAL | 0.5636 | 0.7495 | 0.4514 | 0.0654 | 0.5363 | 0.9421 | 0.1493 | 0.1474 |
| | | STL | 0.3709 | 0.5720 | 0.1965 | 0.0878 | 0.3585 | 0.7213 | 0.0000 | 0.1229 |
| | 2 | RAN | 0.5704 | 0.8101 | 0.4545 | 0.0764 | 0.5474 | 0.9412 | 0.1481 | 0.1393 |
| | | SAL | 0.5746 | 0.7657 | 0.4759 | 0.0721 | 0.5448 | 0.9463 | 0.1000 | 0.1450 |
| | | STL | 0.3939 | 0.5815 | 0.2785 | 0.0807 | 0.3885 | 0.7899 | 0.0615 | 0.1181 |
| | 4 | RAN | 0.5590 | 0.7400 | 0.4291 | 0.0699 | 0.5261 | 0.9518 | 0.1379 | 0.1505 |
| | | SAL | 0.5657 | 0.7749 | 0.4217 | 0.0807 | 0.5385 | 0.9421 | 0.1538 | 0.1518 |
| | | STL | 0.4149 | 0.5671 | 0.2637 | 0.0766 | 0.4097 | 0.8108 | 0.1190 | 0.1227 |
| | 8 | RAN | 0.5672 | 0.7454 | 0.4291 | 0.0799 | 0.5344 | 0.9393 | 0.1311 | 0.1544 |
| | | SAL | 0.5733 | 0.7913 | 0.4667 | 0.0796 | 0.5469 | 0.9562 | 0.1639 | 0.1415 |
| | | STL | 0.4309 | 0.6667 | 0.2861 | 0.0991 | 0.4245 | 0.8169 | 0.1573 | 0.1319 |
| | 16 | RAN | 0.5652 | 0.7597 | 0.3969 | 0.0868 | 0.5387 | 0.9501 | 0.1667 | 0.1489 |
| | | SAL | 0.5682 | 0.7778 | 0.4675 | 0.0738 | 0.5405 | 0.9500 | 0.1967 | 0.1418 |
| | | STL | 0.3972 | 0.6022 | 0.2885 | 0.0670 | 0.3919 | 0.7105 | 0.1649 | 0.1070 |

Table H.1: grouped-nb-twitter-GM1-ALL-ALL-5

| GM1 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | RAN | 0.4629 | 0.6043 | 0.2738 | 0.0849 | 0.3920 | 0.8720 | 0.0000 | 0.2385 |
| | | SAL | 0.4538 | 0.5875 | 0.3167 | 0.0744 | 0.3872 | 0.9129 | 0.0000 | 0.2225 |
| | | STL | 0.5441 | 0.6411 | 0.4658 | 0.0549 | 0.5242 | 0.9282 | 0.0222 | 0.1598 |
| | 1 | RAN | 0.4195 | 0.5598 | 0.2964 | 0.0646 | 0.3892 | 0.9152 | 0.0000 | 0.1589 |
| | | SAL | 0.4184 | 0.5570 | 0.3208 | 0.0647 | 0.3818 | 0.9016 | 0.0000 | 0.1623 |
| | | STL | 0.2090 | 0.3092 | 0.1297 | 0.0534 | 0.1970 | 0.6731 | 0.0000 | 0.1054 |
| | 2 | RAN | 0.4386 | 0.5517 | 0.3314 | 0.0597 | 0.4009 | 0.8797 | 0.0351 | 0.1602 |
| | | SAL | 0.4308 | 0.6005 | 0.3629 | 0.0699 | 0.3938 | 0.9209 | 0.0000 | 0.1563 |
| | | STL | 0.2320 | 0.3820 | 0.1623 | 0.0573 | 0.2274 | 0.6727 | 0.0270 | 0.1162 |
| | 4 | RAN | 0.4290 | 0.5693 | 0.3675 | 0.0556 | 0.3916 | 0.9151 | 0.0370 | 0.1558 |
| | | SAL | 0.4299 | 0.5990 | 0.3043 | 0.0787 | 0.3940 | 0.9150 | 0.0364 | 0.1627 |
| | | STL | 0.2490 | 0.3628 | 0.1997 | 0.0417 | 0.2456 | 0.6429 | 0.0274 | 0.1078 |
| | 8 | RAN | 0.4357 | 0.5521 | 0.3576 | 0.0534 | 0.4005 | 0.8968 | 0.0000 | 0.1604 |
| | | SAL | 0.4276 | 0.5902 | 0.3372 | 0.0742 | 0.3915 | 0.9002 | 0.0000 | 0.1587 |
| | | STL | 0.2632 | 0.3752 | 0.1692 | 0.0642 | 0.2573 | 0.7893 | 0.0000 | 0.1316 |
| | 16 | RAN | 0.4334 | 0.5562 | 0.3275 | 0.0642 | 0.3996 | 0.8664 | 0.0980 | 0.1537 |
| | | SAL | 0.4319 | 0.5588 | 0.3415 | 0.0676 | 0.3957 | 0.8667 | 0.0714 | 0.1558 |
| | | STL | 0.2271 | 0.3321 | 0.1639 | 0.0468 | 0.2241 | 0.6105 | 0.0000 | 0.1015 |

Table H.2: grouped-nb-twitter-GM1-ALL-ALL-10

| GM1 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | RAN | 0.3215 | 0.3906 | 0.2200 | 0.0511 | 0.2534 | 0.8340 | 0.0000 | 0.2091 |
| | | SAL | 0.2995 | 0.4110 | 0.2265 | 0.0685 | 0.2393 | 0.8560 | 0.0000 | 0.2029 |
| | | STL | 0.3861 | 0.4389 | 0.3553 | 0.0316 | 0.3616 | 0.8430 | 0.0000 | 0.1839 |
| | 1 | RAN | 0.2926 | 0.3653 | 0.2279 | 0.0465 | 0.2572 | 0.8185 | 0.0000 | 0.1560 |
| | | SAL | 0.2900 | 0.3541 | 0.2492 | 0.0347 | 0.2553 | 0.7879 | 0.0222 | 0.1510 |
| | | STL | 0.0953 | 0.1352 | 0.0731 | 0.0212 | 0.0895 | 0.5030 | 0.0000 | 0.0860 |
| | 2 | RAN | 0.2942 | 0.3371 | 0.2473 | 0.0266 | 0.2566 | 0.7992 | 0.0000 | 0.1540 |
| | | SAL | 0.2951 | 0.3574 | 0.2445 | 0.0408 | 0.2588 | 0.8385 | 0.0290 | 0.1576 |
| | | STL | 0.1005 | 0.1333 | 0.0817 | 0.0177 | 0.0971 | 0.5591 | 0.0000 | 0.0827 |
| | 4 | RAN | 0.2913 | 0.3444 | 0.2572 | 0.0330 | 0.2530 | 0.8356 | 0.0000 | 0.1567 |
| | | SAL | 0.2964 | 0.3539 | 0.2271 | 0.0448 | 0.2626 | 0.8615 | 0.0000 | 0.1639 |
| | | STL | 0.1119 | 0.1248 | 0.0956 | 0.0101 | 0.1107 | 0.6250 | 0.0000 | 0.0837 |
| | 8 | RAN | 0.2996 | 0.3882 | 0.2327 | 0.0475 | 0.2595 | 0.8249 | 0.0000 | 0.1521 |
| | | SAL | 0.2985 | 0.3701 | 0.2388 | 0.0466 | 0.2615 | 0.8615 | 0.0000 | 0.1629 |
| | | STL | 0.1213 | 0.1617 | 0.0848 | 0.0263 | 0.1186 | 0.6557 | 0.0000 | 0.1046 |
| | 16 | RAN | 0.2987 | 0.3319 | 0.2712 | 0.0229 | 0.2537 | 0.8168 | 0.0274 | 0.1516 |
| | | SAL | 0.2937 | 0.3517 | 0.2543 | 0.0349 | 0.2548 | 0.8803 | 0.0000 | 0.1574 |
| | | STL | 0.0988 | 0.1172 | 0.0733 | 0.0139 | 0.0984 | 0.5714 | 0.0000 | 0.0820 |

Table H.3: grouped-nb-twitter-GM1-ALL-ALL-25

| GM1 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | RAN | 0.2120 | 0.2566 | 0.1395 | 0.0517 | 0.1665 | 0.8214 | 0.0000 | 0.1818 |
| | | SAL | 0.2079 | 0.2373 | 0.1680 | 0.0293 | 0.1610 | 0.8426 | 0.0000 | 0.1748 |
| | | STL | 0.2779 | 0.3191 | 0.2359 | 0.0340 | 0.2485 | 0.8333 | 0.0000 | 0.1849 |
| | 1 | RAN | 0.2158 | 0.2421 | 0.1936 | 0.0200 | 0.1786 | 0.8116 | 0.0000 | 0.1462 |
| | | SAL | 0.2131 | 0.2449 | 0.1897 | 0.0233 | 0.1755 | 0.7368 | 0.0000 | 0.1450 |
| | | STL | 0.0495 | 0.0593 | 0.0387 | 0.0084 | 0.0428 | 0.4078 | 0.0000 | 0.0618 |
| | 2 | RAN | 0.2145 | 0.2501 | 0.1686 | 0.0341 | 0.1768 | 0.8615 | 0.0000 | 0.1393 |
| | | SAL | 0.2186 | 0.2465 | 0.2018 | 0.0199 | 0.1809 | 0.7857 | 0.0000 | 0.1458 |
| | | STL | 0.0563 | 0.0652 | 0.0460 | 0.0079 | 0.0522 | 0.5000 | 0.0000 | 0.0689 |
| | 4 | RAN | 0.2154 | 0.2507 | 0.1683 | 0.0346 | 0.1773 | 0.7924 | 0.0000 | 0.1411 |
| | | SAL | 0.2174 | 0.2467 | 0.1665 | 0.0361 | 0.1818 | 0.8000 | 0.0000 | 0.1444 |
| | | STL | 0.0632 | 0.0746 | 0.0556 | 0.0082 | 0.0621 | 0.5902 | 0.0000 | 0.0716 |
| | 8 | RAN | 0.2195 | 0.2522 | 0.1986 | 0.0234 | 0.1816 | 0.7917 | 0.0000 | 0.1439 |
| | | SAL | 0.2181 | 0.2460 | 0.1960 | 0.0208 | 0.1823 | 0.8750 | 0.0000 | 0.1515 |
| | | STL | 0.0637 | 0.0796 | 0.0504 | 0.0121 | 0.0615 | 0.4940 | 0.0000 | 0.0789 |
| | 16 | RAN | 0.2213 | 0.2436 | 0.1849 | 0.0259 | 0.1845 | 0.7883 | 0.0000 | 0.1489 |
| | | SAL | 0.2183 | 0.2289 | 0.2064 | 0.0092 | 0.1805 | 0.8485 | 0.0000 | 0.1482 |
| | | STL | 0.0520 | 0.0585 | 0.0486 | 0.0046 | 0.0510 | 0.5417 | 0.0000 | 0.0729 |

Table H.4: grouped-nb-twitter-GM1-ALL-ALL-50

| GM1 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | | | F-Score | | | |
| Group Size | Web1T % | Group Type | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | RAN | 0.1644 | 0.1890 | 0.1399 | 0.0246 | 0.1233 | 0.7876 | 0.0000 | 0.1582 |
| | | SAL | 0.1678 | 0.1846 | 0.1509 | 0.0169 | 0.1259 | 0.7807 | 0.0000 | 0.1713 |
| | | STL | 0.2137 | 0.2356 | 0.1917 | 0.0220 | 0.1825 | 0.7818 | 0.0000 | 0.1721 |
| | 1 | RAN | 0.1846 | 0.2014 | 0.1679 | 0.0167 | 0.1487 | 0.7324 | 0.0000 | 0.1373 |
| | | SAL | 0.1785 | 0.1964 | 0.1606 | 0.0179 | 0.1441 | 0.7027 | 0.0000 | 0.1387 |
| | | STL | 0.0351 | 0.0448 | 0.0254 | 0.0097 | 0.0281 | 0.3841 | 0.0000 | 0.0492 |
| | 2 | RAN | 0.1798 | 0.2120 | 0.1475 | 0.0323 | 0.1445 | 0.8000 | 0.0000 | 0.1412 |
| | | SAL | 0.1885 | 0.2167 | 0.1604 | 0.0282 | 0.1537 | 0.7647 | 0.0000 | 0.1404 |
| | | STL | 0.0408 | 0.0512 | 0.0303 | 0.0105 | 0.0362 | 0.3975 | 0.0000 | 0.0572 |
| | 4 | RAN | 0.1855 | 0.1888 | 0.1822 | 0.0033 | 0.1467 | 0.7598 | 0.0000 | 0.1332 |
| | | SAL | 0.1813 | 0.2184 | 0.1442 | 0.0371 | 0.1494 | 0.7941 | 0.0000 | 0.1432 |
| | | STL | 0.0441 | 0.0483 | 0.0400 | 0.0042 | 0.0435 | 0.5357 | 0.0000 | 0.0619 |
| | 8 | RAN | 0.1876 | 0.1927 | 0.1825 | 0.0051 | 0.1529 | 0.7037 | 0.0000 | 0.1392 |
| | | SAL | 0.1813 | 0.2275 | 0.1350 | 0.0462 | 0.1487 | 0.7617 | 0.0000 | 0.1399 |
| | | STL | 0.0481 | 0.0559 | 0.0404 | 0.0077 | 0.0453 | 0.4713 | 0.0000 | 0.0709 |
| | 16 | RAN | 0.1854 | 0.1928 | 0.1779 | 0.0074 | 0.1493 | 0.6950 | 0.0000 | 0.1318 |
| | | SAL | 0.1881 | 0.2114 | 0.1647 | 0.0233 | 0.1509 | 0.7680 | 0.0000 | 0.1382 |
| | | STL | 0.0347 | 0.0379 | 0.0315 | 0.0032 | 0.0324 | 0.3111 | 0.0000 | 0.0503 |

Table H.5: grouped-nb-twitter-GM1-ALL-ALL-75

| GM1 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | RAN | 0.1247 | 0.1247 | 0.1247 | 0.0000 | 0.0843 | 0.7565 | 0.0000 | 0.1440 |
| | | SAL | 0.1254 | 0.1254 | 0.1254 | 0.0000 | 0.0860 | 0.7458 | 0.0000 | 0.1451 |
| | | STL | 0.1254 | 0.1254 | 0.1254 | 0.0000 | 0.0860 | 0.7458 | 0.0000 | 0.1451 |
| | 1 | RAN | 0.1353 | 0.1353 | 0.1353 | 0.0000 | 0.1021 | 0.6933 | 0.0000 | 0.1224 |
| | | SAL | 0.1353 | 0.1353 | 0.1353 | 0.0000 | 0.1042 | 0.7500 | 0.0000 | 0.1246 |
| | | STL | 0.0239 | 0.0239 | 0.0239 | 0.0000 | 0.0172 | 0.3232 | 0.0000 | 0.0410 |
| | 2 | RAN | 0.1344 | 0.1344 | 0.1344 | 0.0000 | 0.1003 | 0.6753 | 0.0000 | 0.1222 |
| | | SAL | 0.1344 | 0.1344 | 0.1344 | 0.0000 | 0.1019 | 0.7027 | 0.0000 | 0.1219 |
| | | STL | 0.0248 | 0.0248 | 0.0248 | 0.0000 | 0.0192 | 0.3333 | 0.0000 | 0.0463 |
| | 4 | RAN | 0.1342 | 0.1342 | 0.1342 | 0.0000 | 0.1028 | 0.6933 | 0.0000 | 0.1212 |
| | | SAL | 0.1351 | 0.1351 | 0.1351 | 0.0000 | 0.1028 | 0.7324 | 0.0000 | 0.1225 |
| | | STL | 0.0231 | 0.0231 | 0.0231 | 0.0000 | 0.0224 | 0.3182 | 0.0000 | 0.0414 |
| | 8 | RAN | 0.1335 | 0.1335 | 0.1335 | 0.0000 | 0.1004 | 0.7027 | 0.0000 | 0.1209 |
| | | SAL | 0.1350 | 0.1350 | 0.1350 | 0.0000 | 0.1029 | 0.7123 | 0.0000 | 0.1220 |
| | | STL | 0.0274 | 0.0274 | 0.0274 | 0.0000 | 0.0230 | 0.4013 | 0.0000 | 0.0522 |
| | 16 | RAN | 0.1386 | 0.1386 | 0.1386 | 0.0000 | 0.1053 | 0.6753 | 0.0000 | 0.1198 |
| | | SAL | 0.1363 | 0.1363 | 0.1363 | 0.0000 | 0.1039 | 0.6923 | 0.0000 | 0.1242 |
| | | STL | 0.0200 | 0.0200 | 0.0200 | 0.0000 | 0.0181 | 0.1991 | 0.0000 | 0.0343 |

Table H.6: grouped-nb-twitter-GM1-ALL-ALL-150

| GM2 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | RAN | 0.5732 | 0.7657 | 0.3986 | 0.0847 | 0.5224 | 0.9486 | 0.0351 | 0.1884 |
| | | SAL | 0.5623 | 0.7890 | 0.3875 | 0.0874 | 0.5087 | 0.9105 | 0.0000 | 0.1956 |
| | | STL | 0.5778 | 0.7356 | 0.4151 | 0.0831 | 0.5585 | 0.9231 | 0.0606 | 0.1564 |
| | 1 | RAN | 0.5874 | 0.7616 | 0.3830 | 0.0780 | 0.5716 | 0.9250 | 0.2222 | 0.1359 |
| | | SAL | 0.5877 | 0.7643 | 0.4103 | 0.0877 | 0.5718 | 0.9153 | 0.2059 | 0.1392 |
| | | STL | 0.4007 | 0.5500 | 0.2871 | 0.0618 | 0.3970 | 0.8182 | 0.1667 | 0.1099 |
| | 2 | RAN | 0.5820 | 0.7467 | 0.4236 | 0.0878 | 0.5651 | 0.9328 | 0.1687 | 0.1425 |
| | | SAL | 0.5844 | 0.7114 | 0.4590 | 0.0751 | 0.5671 | 0.9043 | 0.2078 | 0.1333 |
| | | STL | 0.4240 | 0.5590 | 0.3114 | 0.0733 | 0.4203 | 0.8000 | 0.1284 | 0.1214 |
| | 4 | RAN | 0.5888 | 0.7574 | 0.4727 | 0.0702 | 0.5667 | 0.9237 | 0.2368 | 0.1398 |
| | | SAL | 0.5806 | 0.7926 | 0.4336 | 0.0883 | 0.5599 | 0.9167 | 0.2381 | 0.1499 |
| | | STL | 0.4510 | 0.6162 | 0.3519 | 0.0639 | 0.4511 | 0.8615 | 0.1522 | 0.1193 |
| | 8 | RAN | 0.6026 | 0.7797 | 0.4710 | 0.0839 | 0.5817 | 0.9457 | 0.1944 | 0.1470 |
| | | SAL | 0.6003 | 0.7992 | 0.4743 | 0.0699 | 0.5806 | 0.9237 | 0.2626 | 0.1369 |
| | | STL | 0.4348 | 0.6199 | 0.2906 | 0.0696 | 0.4316 | 0.7950 | 0.2329 | 0.1155 |
| | 16 | RAN | 0.6003 | 0.8437 | 0.4465 | 0.0858 | 0.5808 | 0.9560 | 0.3143 | 0.1318 |
| | | SAL | 0.5884 | 0.7737 | 0.4029 | 0.0999 | 0.5632 | 0.9492 | 0.2118 | 0.1585 |
| | | STL | 0.4340 | 0.5185 | 0.2975 | 0.0667 | 0.4307 | 0.8186 | 0.1481 | 0.1224 |

Table H.7: grouped-nb-twitter-GM2-ALL-ALL-5

| GM2 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | RAN | 0.4438 | 0.6199 | 0.2809 | 0.0829 | 0.3966 | 0.8710 | 0.0000 | 0.1959 |
| | | SAL | 0.4346 | 0.6304 | 0.3645 | 0.0752 | 0.3817 | 0.8429 | 0.0000 | 0.1963 |
| | | STL | 0.4532 | 0.5707 | 0.3270 | 0.0643 | 0.4313 | 0.8824 | 0.0357 | 0.1684 |
| | 1 | RAN | 0.4539 | 0.5810 | 0.3958 | 0.0511 | 0.4332 | 0.8974 | 0.0377 | 0.1542 |
| | | SAL | 0.4511 | 0.5960 | 0.3042 | 0.0716 | 0.4328 | 0.8642 | 0.0702 | 0.1585 |
| | | STL | 0.2489 | 0.3478 | 0.1816 | 0.0377 | 0.2489 | 0.6765 | 0.0513 | 0.1089 |
| | 2 | RAN | 0.4536 | 0.5446 | 0.3374 | 0.0619 | 0.4356 | 0.8988 | 0.0822 | 0.1631 |
| | | SAL | 0.4624 | 0.5246 | 0.4006 | 0.0383 | 0.4402 | 0.8947 | 0.0519 | 0.1549 |
| | | STL | 0.2582 | 0.3644 | 0.1842 | 0.0450 | 0.2586 | 0.7869 | 0.0594 | 0.1230 |
| | 4 | RAN | 0.4538 | 0.5885 | 0.3429 | 0.0805 | 0.4318 | 0.8947 | 0.0759 | 0.1613 |
| | | SAL | 0.4641 | 0.6610 | 0.3768 | 0.0696 | 0.4397 | 0.8848 | 0.0698 | 0.1629 |
| | | STL | 0.2875 | 0.3425 | 0.2146 | 0.0396 | 0.2922 | 0.8116 | 0.0588 | 0.1266 |
| | 8 | RAN | 0.4715 | 0.6051 | 0.3775 | 0.0751 | 0.4482 | 0.9231 | 0.0976 | 0.1628 |
| | | SAL | 0.4696 | 0.5573 | 0.4150 | 0.0468 | 0.4424 | 0.9217 | 0.1290 | 0.1593 |
| | | STL | 0.2727 | 0.3638 | 0.1764 | 0.0402 | 0.2690 | 0.7102 | 0.0250 | 0.1136 |
| | 16 | RAN | 0.4653 | 0.6179 | 0.3557 | 0.0767 | 0.4388 | 0.9043 | 0.0357 | 0.1733 |
| | | SAL | 0.4692 | 0.6117 | 0.3640 | 0.0867 | 0.4367 | 0.9170 | 0.0811 | 0.1741 |
| | | STL | 0.2748 | 0.3317 | 0.1855 | 0.0372 | 0.2742 | 0.7718 | 0.0426 | 0.1210 |

Table H.8: grouped-nb-twitter-GM2-ALL-ALL-10

| GM2 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | RAN | 0.3179 | 0.3833 | 0.2853 | 0.0318 | 0.2706 | 0.8182 | 0.0000 | 0.1887 |
| | | SAL | 0.3253 | 0.3855 | 0.2750 | 0.0415 | 0.2745 | 0.8462 | 0.0000 | 0.1864 |
| | | STL | 0.3212 | 0.3755 | 0.2617 | 0.0401 | 0.3013 | 0.8182 | 0.0000 | 0.1769 |
| | 1 | RAN | 0.3223 | 0.3457 | 0.2971 | 0.0175 | 0.3091 | 0.8571 | 0.0000 | 0.1664 |
| | | SAL | 0.3218 | 0.4249 | 0.2574 | 0.0537 | 0.3045 | 0.8755 | 0.0377 | 0.1658 |
| | | STL | 0.1082 | 0.1254 | 0.0872 | 0.0126 | 0.1083 | 0.5600 | 0.0000 | 0.0857 |
| | 2 | RAN | 0.3176 | 0.3636 | 0.2754 | 0.0346 | 0.3024 | 0.8745 | 0.0645 | 0.1616 |
| | | SAL | 0.3238 | 0.3846 | 0.2684 | 0.0404 | 0.3092 | 0.8308 | 0.0000 | 0.1584 |
| | | STL | 0.1116 | 0.1358 | 0.0963 | 0.0132 | 0.1103 | 0.7368 | 0.0000 | 0.0967 |
| | 4 | RAN | 0.3338 | 0.3673 | 0.3033 | 0.0250 | 0.3104 | 0.8000 | 0.0278 | 0.1652 |
| | | SAL | 0.3317 | 0.4173 | 0.2733 | 0.0458 | 0.3092 | 0.8364 | 0.0270 | 0.1673 |
| | | STL | 0.1404 | 0.1616 | 0.1121 | 0.0162 | 0.1480 | 0.7879 | 0.0000 | 0.1096 |
| | 8 | RAN | 0.3426 | 0.4296 | 0.3064 | 0.0470 | 0.3186 | 0.8681 | 0.0513 | 0.1623 |
| | | SAL | 0.3463 | 0.4602 | 0.2814 | 0.0561 | 0.3200 | 0.8671 | 0.0455 | 0.1730 |
| | | STL | 0.1415 | 0.1684 | 0.1198 | 0.0164 | 0.1404 | 0.6600 | 0.0000 | 0.1023 |
| | 16 | RAN | 0.3351 | 0.3989 | 0.2781 | 0.0353 | 0.3077 | 0.8358 | 0.0299 | 0.1709 |
| | | SAL | 0.3339 | 0.4437 | 0.2787 | 0.0644 | 0.3114 | 0.8619 | 0.0000 | 0.1785 |
| | | STL | 0.1418 | 0.1603 | 0.1105 | 0.0168 | 0.1445 | 0.7048 | 0.0000 | 0.1111 |

Table H.9: grouped-nb-twitter-GM2-ALL-ALL-25

| GM2 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | RAN | 0.2499 | 0.2846 | 0.2213 | 0.0262 | 0.2128 | 0.8438 | 0.0000 | 0.1727 |
| | | SAL | 0.2531 | 0.2662 | 0.2349 | 0.0133 | 0.2069 | 0.7941 | 0.0000 | 0.1716 |
| | | STL | 0.2497 | 0.2992 | 0.2038 | 0.0390 | 0.2328 | 0.7826 | 0.0000 | 0.1661 |
| | 1 | RAN | 0.2522 | 0.3044 | 0.2198 | 0.0373 | 0.2375 | 0.7753 | 0.0000 | 0.1531 |
| | | SAL | 0.2467 | 0.2626 | 0.2193 | 0.0194 | 0.2326 | 0.8288 | 0.0000 | 0.1600 |
| | | STL | 0.0573 | 0.0611 | 0.0507 | 0.0047 | 0.0573 | 0.5106 | 0.0000 | 0.0718 |
| | 2 | RAN | 0.2477 | 0.2533 | 0.2367 | 0.0077 | 0.2329 | 0.8189 | 0.0000 | 0.1636 |
| | | SAL | 0.2481 | 0.2660 | 0.2241 | 0.0177 | 0.2351 | 0.8300 | 0.0000 | 0.1627 |
| | | STL | 0.0571 | 0.0730 | 0.0413 | 0.0129 | 0.0550 | 0.4783 | 0.0000 | 0.0707 |
| | 4 | RAN | 0.2583 | 0.2667 | 0.2498 | 0.0069 | 0.2344 | 0.7944 | 0.0000 | 0.1647 |
| | | SAL | 0.2617 | 0.3055 | 0.2335 | 0.0314 | 0.2398 | 0.8037 | 0.0420 | 0.1618 |
| | | STL | 0.0739 | 0.0946 | 0.0578 | 0.0154 | 0.0789 | 0.7419 | 0.0000 | 0.0904 |
| | 8 | RAN | 0.2674 | 0.2880 | 0.2415 | 0.0194 | 0.2451 | 0.8117 | 0.0250 | 0.1623 |
| | | SAL | 0.2678 | 0.2806 | 0.2607 | 0.0090 | 0.2465 | 0.8362 | 0.0000 | 0.1669 |
| | | STL | 0.0822 | 0.0891 | 0.0771 | 0.0051 | 0.0840 | 0.5241 | 0.0000 | 0.0836 |
| | 16 | RAN | 0.2629 | 0.3031 | 0.2405 | 0.0285 | 0.2408 | 0.8438 | 0.0204 | 0.1690 |
| | | SAL | 0.2624 | 0.2887 | 0.2138 | 0.0344 | 0.2419 | 0.8313 | 0.0000 | 0.1745 |
| | | STL | 0.0817 | 0.0870 | 0.0786 | 0.0038 | 0.0846 | 0.5310 | 0.0000 | 0.0866 |

Table H.10: grouped-nb-twitter-GM2-ALL-ALL-50

| GM2 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | RAN | 0.2125 | 0.2621 | 0.1630 | 0.0496 | 0.1744 | 0.7385 | 0.0000 | 0.1610 |
| | | SAL | 0.2195 | 0.2291 | 0.2100 | 0.0095 | 0.1828 | 0.7397 | 0.0000 | 0.1634 |
| | | STL | 0.2164 | 0.2458 | 0.1871 | 0.0293 | 0.1965 | 0.7879 | 0.0000 | 0.1614 |
| | 1 | RAN | 0.2051 | 0.2181 | 0.1922 | 0.0129 | 0.1938 | 0.8000 | 0.0000 | 0.1548 |
| | | SAL | 0.2126 | 0.2358 | 0.1895 | 0.0232 | 0.2000 | 0.8036 | 0.0000 | 0.1578 |
| | | STL | 0.0375 | 0.0382 | 0.0368 | 0.0007 | 0.0356 | 0.3158 | 0.0000 | 0.0500 |
| | 2 | RAN | 0.2131 | 0.2229 | 0.2034 | 0.0098 | 0.2012 | 0.8000 | 0.0000 | 0.1539 |
| | | SAL | 0.2104 | 0.2255 | 0.1953 | 0.0151 | 0.1989 | 0.8293 | 0.0000 | 0.1513 |
| | | STL | 0.0370 | 0.0413 | 0.0328 | 0.0043 | 0.0344 | 0.3721 | 0.0000 | 0.0516 |
| | 4 | RAN | 0.2255 | 0.2477 | 0.2033 | 0.0222 | 0.2067 | 0.8073 | 0.0000 | 0.1553 |
| | | SAL | 0.2244 | 0.2291 | 0.2197 | 0.0047 | 0.2044 | 0.7500 | 0.0000 | 0.1582 |
| | | STL | 0.0501 | 0.0524 | 0.0477 | 0.0024 | 0.0540 | 0.5000 | 0.0000 | 0.0671 |
| | 8 | RAN | 0.2342 | 0.2535 | 0.2149 | 0.0193 | 0.2152 | 0.7887 | 0.0000 | 0.1659 |
| | | SAL | 0.2346 | 0.2503 | 0.2189 | 0.0157 | 0.2144 | 0.8106 | 0.0000 | 0.1664 |
| | | STL | 0.0622 | 0.0666 | 0.0578 | 0.0044 | 0.0648 | 0.5275 | 0.0000 | 0.0788 |
| | 16 | RAN | 0.2253 | 0.2316 | 0.2190 | 0.0063 | 0.2031 | 0.8182 | 0.0000 | 0.1665 |
| | | SAL | 0.2257 | 0.2579 | 0.1935 | 0.0322 | 0.2079 | 0.8571 | 0.0000 | 0.1675 |
| | | STL | 0.0590 | 0.0607 | 0.0573 | 0.0017 | 0.0609 | 0.4786 | 0.0000 | 0.0716 |

Table H.11: grouped-nb-twitter-GM2-ALL-ALL-75

| GM2 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | RAN | 0.1690 | 0.1690 | 0.1690 | 0.0000 | 0.1355 | 0.7619 | 0.0000 | 0.1476 |
| | | SAL | 0.1719 | 0.1719 | 0.1719 | 0.0000 | 0.1390 | 0.7500 | 0.0000 | 0.1488 |
| | | STL | 0.1719 | 0.1719 | 0.1719 | 0.0000 | 0.1390 | 0.7500 | 0.0000 | 0.1488 |
| | 1 | RAN | 0.1621 | 0.1621 | 0.1621 | 0.0000 | 0.1521 | 0.7593 | 0.0000 | 0.1450 |
| | | SAL | 0.1655 | 0.1655 | 0.1655 | 0.0000 | 0.1541 | 0.7321 | 0.0000 | 0.1426 |
| | | STL | 0.0190 | 0.0190 | 0.0190 | 0.0000 | 0.0160 | 0.2629 | 0.0000 | 0.0336 |
| | 2 | RAN | 0.1659 | 0.1659 | 0.1659 | 0.0000 | 0.1565 | 0.7748 | 0.0000 | 0.1487 |
| | | SAL | 0.1655 | 0.1655 | 0.1655 | 0.0000 | 0.1580 | 0.7967 | 0.0000 | 0.1488 |
| | | STL | 0.0206 | 0.0206 | 0.0206 | 0.0000 | 0.0169 | 0.2623 | 0.0000 | 0.0337 |
| | 4 | RAN | 0.1789 | 0.1789 | 0.1789 | 0.0000 | 0.1611 | 0.7477 | 0.0000 | 0.1467 |
| | | SAL | 0.1781 | 0.1781 | 0.1781 | 0.0000 | 0.1611 | 0.7455 | 0.0000 | 0.1481 |
| | | STL | 0.0276 | 0.0276 | 0.0276 | 0.0000 | 0.0286 | 0.3750 | 0.0000 | 0.0491 |
| | 8 | RAN | 0.1810 | 0.1810 | 0.1810 | 0.0000 | 0.1658 | 0.7414 | 0.0000 | 0.1522 |
| | | SAL | 0.1819 | 0.1819 | 0.1819 | 0.0000 | 0.1662 | 0.7339 | 0.0000 | 0.1541 |
| | | STL | 0.0363 | 0.0363 | 0.0363 | 0.0000 | 0.0388 | 0.4483 | 0.0000 | 0.0648 |
| | 16 | RAN | 0.1753 | 0.1753 | 0.1753 | 0.0000 | 0.1570 | 0.7293 | 0.0000 | 0.1544 |
| | | SAL | 0.1760 | 0.1760 | 0.1760 | 0.0000 | 0.1589 | 0.7519 | 0.0000 | 0.1543 |
| | | STL | 0.0357 | 0.0357 | 0.0357 | 0.0000 | 0.0348 | 0.3692 | 0.0000 | 0.0559 |

Table H.12: grouped-nb-twitter-GM2-ALL-ALL-150

| GM5 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | RAN | 0.3442 | 0.4815 | 0.2431 | 0.0621 | 0.2258 | 0.7931 | 0.0000 | 0.1935 |
| | | SAL | 0.3544 | 0.5293 | 0.2909 | 0.0486 | 0.2280 | 0.7931 | 0.0000 | 0.1947 |
| | | STL | 0.3374 | 0.4555 | 0.2564 | 0.0456 | 0.2380 | 0.7719 | 0.0000 | 0.1905 |
| | 1 | RAN | 0.5553 | 0.7495 | 0.4139 | 0.0861 | 0.5284 | 0.8703 | 0.1096 | 0.1466 |
| | | SAL | 0.5531 | 0.7345 | 0.4227 | 0.0845 | 0.5285 | 0.9106 | 0.1379 | 0.1473 |
| | | STL | 0.5506 | 0.6907 | 0.4066 | 0.0661 | 0.5404 | 0.8750 | 0.2330 | 0.1296 |
| | 2 | RAN | 0.5544 | 0.7220 | 0.4161 | 0.0765 | 0.5230 | 0.9206 | 0.1975 | 0.1587 |
| | | SAL | 0.5492 | 0.6859 | 0.4000 | 0.0843 | 0.5206 | 0.9231 | 0.1212 | 0.1572 |
| | | STL | 0.5534 | 0.7087 | 0.4254 | 0.0755 | 0.5403 | 0.8629 | 0.1975 | 0.1417 |
| | 4 | RAN | 0.5595 | 0.7021 | 0.4539 | 0.0604 | 0.5220 | 0.9021 | 0.0426 | 0.1629 |
| | | SAL | 0.5396 | 0.7077 | 0.3814 | 0.0809 | 0.5031 | 0.9052 | 0.0588 | 0.1619 |
| | | STL | 0.5558 | 0.6981 | 0.4280 | 0.0698 | 0.5420 | 0.8689 | 0.2162 | 0.1409 |
| | 8 | RAN | 0.5533 | 0.7071 | 0.3968 | 0.0677 | 0.5163 | 0.9224 | 0.0779 | 0.1625 |
| | | SAL | 0.5432 | 0.6912 | 0.4127 | 0.0717 | 0.5057 | 0.8972 | 0.1067 | 0.1596 |
| | | STL | 0.5686 | 0.7094 | 0.4029 | 0.0642 | 0.5566 | 0.9254 | 0.2338 | 0.1398 |
| | 16 | RAN | 0.5562 | 0.7680 | 0.4106 | 0.0869 | 0.5202 | 0.8713 | 0.0870 | 0.1730 |
| | | SAL | 0.5416 | 0.6900 | 0.4373 | 0.0803 | 0.5070 | 0.9245 | 0.0571 | 0.1704 |
| | | STL | 0.5758 | 0.6852 | 0.4322 | 0.0594 | 0.5626 | 0.8833 | 0.1690 | 0.1425 |

Table H.13: grouped-nb-twitter-GM5-ALL-ALL-5

| GM5 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | RAN | 0.2462 | 0.3357 | 0.2000 | 0.0395 | 0.1752 | 0.8136 | 0.0000 | 0.1570 |
| | | SAL | 0.2450 | 0.3185 | 0.1748 | 0.0312 | 0.1683 | 0.7931 | 0.0000 | 0.1598 |
| | | STL | 0.2313 | 0.3245 | 0.1759 | 0.0331 | 0.1700 | 0.7719 | 0.0000 | 0.1617 |
| | 1 | RAN | 0.4266 | 0.5466 | 0.3396 | 0.0669 | 0.3998 | 0.8175 | 0.0556 | 0.1588 |
| | | SAL | 0.4247 | 0.5410 | 0.3065 | 0.0655 | 0.3970 | 0.8571 | 0.0833 | 0.1555 |
| | | STL | 0.4151 | 0.5362 | 0.3212 | 0.0552 | 0.4015 | 0.8400 | 0.0976 | 0.1490 |
| | 2 | RAN | 0.4105 | 0.5806 | 0.3103 | 0.0674 | 0.3807 | 0.8571 | 0.0811 | 0.1638 |
| | | SAL | 0.4215 | 0.5522 | 0.3041 | 0.0613 | 0.3940 | 0.8889 | 0.0267 | 0.1706 |
| | | STL | 0.4240 | 0.5371 | 0.3035 | 0.0631 | 0.4115 | 0.8618 | 0.1167 | 0.1596 |
| | 4 | RAN | 0.4182 | 0.5954 | 0.3220 | 0.0724 | 0.3833 | 0.8618 | 0.0635 | 0.1716 |
| | | SAL | 0.4140 | 0.4774 | 0.3069 | 0.0481 | 0.3792 | 0.8468 | 0.0286 | 0.1768 |
| | | STL | 0.4252 | 0.5124 | 0.3368 | 0.0531 | 0.4114 | 0.8583 | 0.1096 | 0.1629 |
| | 8 | RAN | 0.4235 | 0.5239 | 0.3269 | 0.0573 | 0.3906 | 0.8367 | 0.0656 | 0.1687 |
| | | SAL | 0.4108 | 0.4825 | 0.3302 | 0.0447 | 0.3796 | 0.8515 | 0.0303 | 0.1709 |
| | | STL | 0.4283 | 0.5090 | 0.3364 | 0.0474 | 0.4156 | 0.8750 | 0.1429 | 0.1601 |
| | 16 | RAN | 0.4333 | 0.5369 | 0.3519 | 0.0559 | 0.3979 | 0.8571 | 0.0278 | 0.1758 |
| | | SAL | 0.4207 | 0.5606 | 0.3228 | 0.0716 | 0.3916 | 0.8932 | 0.0267 | 0.1750 |
| | | STL | 0.4418 | 0.5221 | 0.3723 | 0.0446 | 0.4291 | 0.8480 | 0.1067 | 0.1592 |

Table H.14: grouped-nb-twitter-GM5-ALL-ALL-10

| GM5 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | RAN | 0.1534 | 0.1780 | 0.1256 | 0.0193 | 0.1293 | 0.7931 | 0.0000 | 0.1450 |
| | | SAL | 0.1580 | 0.2141 | 0.1224 | 0.0292 | 0.1265 | 0.7931 | 0.0000 | 0.1413 |
| | | STL | 0.1516 | 0.1929 | 0.1335 | 0.0192 | 0.1307 | 0.7719 | 0.0000 | 0.1464 |
| | 1 | RAN | 0.2966 | 0.3293 | 0.2781 | 0.0196 | 0.2711 | 0.7500 | 0.0000 | 0.1594 |
| | | SAL | 0.3041 | 0.3472 | 0.2514 | 0.0357 | 0.2775 | 0.7143 | 0.0000 | 0.1652 |
| | | STL | 0.2962 | 0.3613 | 0.2703 | 0.0322 | 0.2824 | 0.8190 | 0.0357 | 0.1592 |
| | 2 | RAN | 0.2945 | 0.3486 | 0.2405 | 0.0335 | 0.2728 | 0.8125 | 0.0000 | 0.1682 |
| | | SAL | 0.2923 | 0.3478 | 0.2644 | 0.0319 | 0.2684 | 0.8710 | 0.0000 | 0.1646 |
| | | STL | 0.2901 | 0.3493 | 0.2411 | 0.0317 | 0.2772 | 0.8571 | 0.0278 | 0.1654 |
| | 4 | RAN | 0.2924 | 0.3272 | 0.2133 | 0.0394 | 0.2601 | 0.8750 | 0.0000 | 0.1673 |
| | | SAL | 0.2909 | 0.3271 | 0.2491 | 0.0297 | 0.2614 | 0.8710 | 0.0000 | 0.1761 |
| | | STL | 0.2972 | 0.3300 | 0.2718 | 0.0187 | 0.2851 | 0.8571 | 0.0357 | 0.1712 |
| | 8 | RAN | 0.2914 | 0.3552 | 0.2278 | 0.0417 | 0.2652 | 0.8438 | 0.0000 | 0.1689 |
| | | SAL | 0.2885 | 0.3166 | 0.2632 | 0.0203 | 0.2655 | 0.8125 | 0.0000 | 0.1721 |
| | | STL | 0.2987 | 0.3288 | 0.2746 | 0.0194 | 0.2860 | 0.8387 | 0.0308 | 0.1701 |
| | 16 | RAN | 0.2906 | 0.3765 | 0.2488 | 0.0417 | 0.2659 | 0.8615 | 0.0000 | 0.1761 |
| | | SAL | 0.2946 | 0.3568 | 0.2032 | 0.0506 | 0.2694 | 0.8254 | 0.0000 | 0.1723 |
| | | STL | 0.3068 | 0.3309 | 0.2810 | 0.0154 | 0.2953 | 0.8438 | 0.0286 | 0.1721 |

Table H.15: grouped-nb-twitter-GM5-ALL-ALL-25

| GM5 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | RAN | 0.1269 | 0.1351 | 0.1218 | 0.0059 | 0.1106 | 0.7719 | 0.0000 | 0.1421 |
| | | SAL | 0.1247 | 0.1558 | 0.0999 | 0.0232 | 0.1081 | 0.6909 | 0.0000 | 0.1359 |
| | | STL | 0.1162 | 0.1298 | 0.1065 | 0.0099 | 0.1052 | 0.7241 | 0.0000 | 0.1390 |
| | 1 | RAN | 0.2254 | 0.2481 | 0.1808 | 0.0315 | 0.2063 | 0.7200 | 0.0000 | 0.1517 |
| | | SAL | 0.2287 | 0.2534 | 0.1825 | 0.0327 | 0.2067 | 0.7071 | 0.0000 | 0.1527 |
| | | STL | 0.2289 | 0.2484 | 0.2129 | 0.0147 | 0.2159 | 0.7961 | 0.0000 | 0.1538 |
| | 2 | RAN | 0.2268 | 0.2327 | 0.2218 | 0.0045 | 0.2069 | 0.7647 | 0.0000 | 0.1617 |
| | | SAL | 0.2242 | 0.2637 | 0.2009 | 0.0281 | 0.2027 | 0.8000 | 0.0000 | 0.1624 |
| | | STL | 0.2246 | 0.2511 | 0.1996 | 0.0210 | 0.2156 | 0.8000 | 0.0132 | 0.1645 |
| | 4 | RAN | 0.2215 | 0.2282 | 0.2133 | 0.0062 | 0.1986 | 0.8065 | 0.0000 | 0.1654 |
| | | SAL | 0.2170 | 0.2441 | 0.1770 | 0.0288 | 0.1953 | 0.7937 | 0.0000 | 0.1634 |
| | | STL | 0.2263 | 0.2336 | 0.2141 | 0.0087 | 0.2153 | 0.8254 | 0.0000 | 0.1678 |
| | 8 | RAN | 0.2212 | 0.2573 | 0.2010 | 0.0256 | 0.1959 | 0.7353 | 0.0000 | 0.1593 |
| | | SAL | 0.2160 | 0.2452 | 0.1871 | 0.0237 | 0.1992 | 0.8065 | 0.0000 | 0.1649 |
| | | STL | 0.2280 | 0.2374 | 0.2198 | 0.0073 | 0.2148 | 0.8224 | 0.0000 | 0.1662 |
| | 16 | RAN | 0.2187 | 0.2328 | 0.1954 | 0.0166 | 0.1973 | 0.8710 | 0.0000 | 0.1596 |
| | | SAL | 0.2240 | 0.2608 | 0.1540 | 0.0495 | 0.2010 | 0.8254 | 0.0000 | 0.1653 |
| | | STL | 0.2308 | 0.2400 | 0.2226 | 0.0071 | 0.2194 | 0.8182 | 0.0000 | 0.1674 |

Table H.16: grouped-nb-twitter-GM5-ALL-ALL-50

| GM5 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | RAN | 0.0974 | 0.1049 | 0.0898 | 0.0075 | 0.0914 | 0.7241 | 0.0000 | 0.1239 |
| | | SAL | 0.0942 | 0.1047 | 0.0836 | 0.0106 | 0.0875 | 0.7719 | 0.0000 | 0.1280 |
| | | STL | 0.1014 | 0.1069 | 0.0960 | 0.0054 | 0.0935 | 0.7273 | 0.0000 | 0.1318 |
| | 1 | RAN | 0.1964 | 0.2086 | 0.1843 | 0.0122 | 0.1753 | 0.6588 | 0.0000 | 0.1446 |
| | | SAL | 0.1970 | 0.2006 | 0.1934 | 0.0036 | 0.1803 | 0.6735 | 0.0000 | 0.1471 |
| | | STL | 0.1946 | 0.2111 | 0.1780 | 0.0166 | 0.1790 | 0.6739 | 0.0000 | 0.1454 |
| | 2 | RAN | 0.1957 | 0.2027 | 0.1887 | 0.0070 | 0.1734 | 0.7647 | 0.0000 | 0.1476 |
| | | SAL | 0.1934 | 0.2066 | 0.1803 | 0.0132 | 0.1750 | 0.8710 | 0.0000 | 0.1553 |
| | | STL | 0.1966 | 0.2113 | 0.1819 | 0.0147 | 0.1818 | 0.8254 | 0.0000 | 0.1529 |
| | 4 | RAN | 0.1881 | 0.1984 | 0.1777 | 0.0103 | 0.1705 | 0.7619 | 0.0000 | 0.1595 |
| | | SAL | 0.1860 | 0.1880 | 0.1839 | 0.0021 | 0.1709 | 0.7813 | 0.0000 | 0.1564 |
| | | STL | 0.1895 | 0.1945 | 0.1844 | 0.0051 | 0.1760 | 0.7937 | 0.0000 | 0.1533 |
| | 8 | RAN | 0.1897 | 0.1904 | 0.1890 | 0.0007 | 0.1679 | 0.7576 | 0.0000 | 0.1559 |
| | | SAL | 0.1871 | 0.1880 | 0.1861 | 0.0009 | 0.1701 | 0.7813 | 0.0000 | 0.1553 |
| | | STL | 0.1939 | 0.1945 | 0.1932 | 0.0007 | 0.1806 | 0.8065 | 0.0000 | 0.1581 |
| | 16 | RAN | 0.1858 | 0.2027 | 0.1689 | 0.0169 | 0.1673 | 0.7937 | 0.0000 | 0.1581 |
| | | SAL | 0.1881 | 0.1987 | 0.1775 | 0.0106 | 0.1699 | 0.7500 | 0.0000 | 0.1590 |
| | | STL | 0.1985 | 0.2060 | 0.1910 | 0.0075 | 0.1852 | 0.7937 | 0.0000 | 0.1579 |

Table H.17: grouped-nb-twitter-GM5-ALL-ALL-75

| GM5 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | RAN | 0.0792 | 0.0792 | 0.0792 | 0.0000 | 0.0765 | 0.6429 | 0.0000 | 0.1203 |
| | | SAL | 0.0784 | 0.0784 | 0.0784 | 0.0000 | 0.0749 | 0.6182 | 0.0000 | 0.1178 |
| | | STL | 0.0784 | 0.0784 | 0.0784 | 0.0000 | 0.0749 | 0.6182 | 0.0000 | 0.1178 |
| | 1 | RAN | 0.1542 | 0.1542 | 0.1542 | 0.0000 | 0.1380 | 0.6392 | 0.0000 | 0.1370 |
| | | SAL | 0.1531 | 0.1531 | 0.1531 | 0.0000 | 0.1364 | 0.6263 | 0.0000 | 0.1354 |
| | | STL | 0.1531 | 0.1531 | 0.1531 | 0.0000 | 0.1364 | 0.6263 | 0.0000 | 0.1354 |
| | 2 | RAN | 0.1531 | 0.1531 | 0.1531 | 0.0000 | 0.1363 | 0.7692 | 0.0000 | 0.1425 |
| | | SAL | 0.1520 | 0.1520 | 0.1520 | 0.0000 | 0.1358 | 0.7879 | 0.0000 | 0.1428 |
| | | STL | 0.1520 | 0.1520 | 0.1520 | 0.0000 | 0.1358 | 0.7879 | 0.0000 | 0.1428 |
| | 4 | RAN | 0.1414 | 0.1414 | 0.1414 | 0.0000 | 0.1295 | 0.7500 | 0.0000 | 0.1463 |
| | | SAL | 0.1399 | 0.1399 | 0.1399 | 0.0000 | 0.1266 | 0.7500 | 0.0000 | 0.1456 |
| | | STL | 0.1399 | 0.1399 | 0.1399 | 0.0000 | 0.1266 | 0.7500 | 0.0000 | 0.1456 |
| | 8 | RAN | 0.1404 | 0.1404 | 0.1404 | 0.0000 | 0.1268 | 0.7463 | 0.0000 | 0.1457 |
| | | SAL | 0.1399 | 0.1399 | 0.1399 | 0.0000 | 0.1267 | 0.7302 | 0.0000 | 0.1436 |
| | | STL | 0.1399 | 0.1399 | 0.1399 | 0.0000 | 0.1267 | 0.7302 | 0.0000 | 0.1436 |
| | 16 | RAN | 0.1430 | 0.1430 | 0.1430 | 0.0000 | 0.1292 | 0.7500 | 0.0000 | 0.1467 |
| | | SAL | 0.1420 | 0.1420 | 0.1420 | 0.0000 | 0.1295 | 0.7813 | 0.0000 | 0.1464 |
| | | STL | 0.1420 | 0.1420 | 0.1420 | 0.0000 | 0.1295 | 0.7813 | 0.0000 | 0.1464 |

Table H.18: grouped-nb-twitter-GM5-ALL-ALL-150

| GB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | RAN | 0.6189 | 0.8011 | 0.4860 | 0.0817 | 0.5761 | 0.9234 | 0.0392 | 0.1733 |
| | | SAL | 0.6135 | 0.8021 | 0.5154 | 0.0624 | 0.5689 | 0.9284 | 0.0800 | 0.1732 |
| | | STL | 0.6215 | 0.7500 | 0.4745 | 0.0731 | 0.5991 | 0.9112 | 0.1231 | 0.1466 |
| | 1 | RAN | 0.6132 | 0.8125 | 0.4816 | 0.0678 | 0.5976 | 0.9480 | 0.2051 | 0.1291 |
| | | SAL | 0.6155 | 0.7584 | 0.4585 | 0.0765 | 0.5967 | 0.9118 | 0.2000 | 0.1344 |
| | | STL | 0.6040 | 0.7840 | 0.4606 | 0.0752 | 0.5933 | 0.9126 | 0.2783 | 0.1250 |
| | 2 | RAN | 0.6183 | 0.7311 | 0.4954 | 0.0638 | 0.5982 | 0.9347 | 0.3077 | 0.1280 |
| | | SAL | 0.6223 | 0.7823 | 0.5029 | 0.0766 | 0.5998 | 0.9130 | 0.2340 | 0.1377 |
| | | STL | 0.6075 | 0.7653 | 0.4606 | 0.0678 | 0.5978 | 0.9228 | 0.2824 | 0.1210 |
| | 4 | RAN | 0.6257 | 0.8546 | 0.4694 | 0.0850 | 0.6075 | 0.9613 | 0.2444 | 0.1387 |
| | | SAL | 0.6291 | 0.8187 | 0.5164 | 0.0874 | 0.6085 | 0.9320 | 0.2500 | 0.1387 |
| | | STL | 0.6048 | 0.7781 | 0.4953 | 0.0700 | 0.5969 | 0.9208 | 0.3218 | 0.1204 |
| | 8 | RAN | 0.6210 | 0.8474 | 0.4717 | 0.0767 | 0.6035 | 0.9409 | 0.2500 | 0.1318 |
| | | SAL | 0.6247 | 0.8013 | 0.4734 | 0.0831 | 0.6055 | 0.9203 | 0.2222 | 0.1310 |
| | | STL | 0.6106 | 0.7813 | 0.4669 | 0.0699 | 0.6029 | 0.9208 | 0.2459 | 0.1164 |
| | 16 | RAN | 0.6345 | 0.8216 | 0.4648 | 0.0810 | 0.6169 | 0.9512 | 0.2299 | 0.1355 |
| | | SAL | 0.6346 | 0.7564 | 0.5229 | 0.0711 | 0.6104 | 0.9070 | 0.2154 | 0.1368 |
| | | STL | 0.6106 | 0.7867 | 0.4795 | 0.0770 | 0.6021 | 0.9208 | 0.2526 | 0.1235 |

Table H.19: grouped-nb-twitter-GB3-ALL-ALL-5

| GB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | RAN | 0.5006 | 0.6108 | 0.3844 | 0.0595 | 0.4537 | 0.8696 | 0.0400 | 0.1853 |
| | | SAL | 0.4898 | 0.5952 | 0.3958 | 0.0491 | 0.4434 | 0.9091 | 0.0000 | 0.1827 |
| | | STL | 0.4940 | 0.6108 | 0.3477 | 0.0709 | 0.4658 | 0.8817 | 0.0000 | 0.1708 |
| | 1 | RAN | 0.5038 | 0.6000 | 0.4232 | 0.0499 | 0.4769 | 0.8871 | 0.1333 | 0.1549 |
| | | SAL | 0.4977 | 0.6266 | 0.3297 | 0.0742 | 0.4743 | 0.8745 | 0.1190 | 0.1562 |
| | | STL | 0.4909 | 0.6356 | 0.3796 | 0.0661 | 0.4769 | 0.9002 | 0.1905 | 0.1430 |
| | 2 | RAN | 0.5092 | 0.7366 | 0.4142 | 0.0746 | 0.4865 | 0.9289 | 0.1481 | 0.1447 |
| | | SAL | 0.5019 | 0.6596 | 0.3845 | 0.0773 | 0.4754 | 0.8716 | 0.1075 | 0.1510 |
| | | STL | 0.4935 | 0.6094 | 0.3748 | 0.0594 | 0.4809 | 0.9031 | 0.1647 | 0.1371 |
| | 4 | RAN | 0.5100 | 0.6671 | 0.4123 | 0.0825 | 0.4850 | 0.9066 | 0.0741 | 0.1637 |
| | | SAL | 0.5039 | 0.6517 | 0.3665 | 0.0703 | 0.4782 | 0.9102 | 0.0899 | 0.1496 |
| | | STL | 0.4895 | 0.6067 | 0.3939 | 0.0608 | 0.4770 | 0.8994 | 0.1649 | 0.1419 |
| | 8 | RAN | 0.5061 | 0.6355 | 0.4196 | 0.0675 | 0.4813 | 0.9197 | 0.1389 | 0.1561 |
| | | SAL | 0.5013 | 0.6359 | 0.3893 | 0.0788 | 0.4774 | 0.8854 | 0.0952 | 0.1534 |
| | | STL | 0.4924 | 0.6218 | 0.3732 | 0.0671 | 0.4818 | 0.8994 | 0.1831 | 0.1363 |
| | 16 | RAN | 0.5038 | 0.6205 | 0.3986 | 0.0662 | 0.4811 | 0.8696 | 0.0909 | 0.1464 |
| | | SAL | 0.5098 | 0.6137 | 0.3997 | 0.0550 | 0.4842 | 0.8986 | 0.0941 | 0.1517 |
| | | STL | 0.4987 | 0.6385 | 0.3876 | 0.0682 | 0.4861 | 0.8975 | 0.1778 | 0.1412 |

Table H.20: grouped-nb-twitter-GB3-ALL-ALL-10

| GB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | RAN | 0.3602 | 0.4185 | 0.3076 | 0.0371 | 0.3105 | 0.8000 | 0.0000 | 0.1891 |
| | | SAL | 0.3614 | 0.4401 | 0.3153 | 0.0393 | 0.3135 | 0.8696 | 0.0000 | 0.1888 |
| | | STL | 0.3535 | 0.4211 | 0.2988 | 0.0421 | 0.3253 | 0.7904 | 0.0000 | 0.1799 |
| | 1 | RAN | 0.3788 | 0.4236 | 0.3384 | 0.0317 | 0.3517 | 0.8473 | 0.0256 | 0.1581 |
| | | SAL | 0.3764 | 0.4429 | 0.3147 | 0.0411 | 0.3518 | 0.8824 | 0.0202 | 0.1576 |
| | | STL | 0.3619 | 0.4222 | 0.3190 | 0.0344 | 0.3436 | 0.8393 | 0.0449 | 0.1511 |
| | 2 | RAN | 0.3817 | 0.4394 | 0.3367 | 0.0397 | 0.3563 | 0.7629 | 0.0253 | 0.1525 |
| | | SAL | 0.3796 | 0.4232 | 0.3464 | 0.0266 | 0.3512 | 0.7858 | 0.0227 | 0.1501 |
| | | STL | 0.3667 | 0.4361 | 0.3271 | 0.0360 | 0.3516 | 0.8155 | 0.0440 | 0.1493 |
| | 4 | RAN | 0.3882 | 0.4834 | 0.3428 | 0.0447 | 0.3599 | 0.8504 | 0.0294 | 0.1609 |
| | | SAL | 0.3832 | 0.4562 | 0.3207 | 0.0522 | 0.3561 | 0.7828 | 0.0227 | 0.1530 |
| | | STL | 0.3604 | 0.4353 | 0.3245 | 0.0369 | 0.3457 | 0.8121 | 0.0440 | 0.1519 |
| | 8 | RAN | 0.3733 | 0.4633 | 0.3186 | 0.0511 | 0.3512 | 0.8355 | 0.0588 | 0.1528 |
| | | SAL | 0.3824 | 0.4558 | 0.3353 | 0.0409 | 0.3565 | 0.7815 | 0.0244 | 0.1536 |
| | | STL | 0.3642 | 0.4361 | 0.3206 | 0.0394 | 0.3508 | 0.8117 | 0.0842 | 0.1506 |
| | 16 | RAN | 0.3868 | 0.4290 | 0.2890 | 0.0469 | 0.3595 | 0.8649 | 0.0400 | 0.1685 |
| | | SAL | 0.3915 | 0.4422 | 0.3474 | 0.0357 | 0.3625 | 0.8451 | 0.0270 | 0.1651 |
| | | STL | 0.3730 | 0.4431 | 0.3184 | 0.0395 | 0.3580 | 0.8493 | 0.0449 | 0.1515 |

Table H.21: grouped-nb-twitter-GB3-ALL-ALL-25

| GB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | RAN | 0.2849 | 0.3171 | 0.2642 | 0.0231 | 0.2400 | 0.7606 | 0.0000 | 0.1753 |
| | | SAL | 0.2831 | 0.3015 | 0.2612 | 0.0166 | 0.2401 | 0.8955 | 0.0000 | 0.1791 |
| | | STL | 0.2837 | 0.3292 | 0.2411 | 0.0360 | 0.2569 | 0.7788 | 0.0000 | 0.1759 |
| | 1 | RAN | 0.3022 | 0.3356 | 0.2757 | 0.0249 | 0.2751 | 0.7892 | 0.0000 | 0.1555 |
| | | SAL | 0.3002 | 0.3108 | 0.2923 | 0.0078 | 0.2742 | 0.8219 | 0.0235 | 0.1583 |
| | | STL | 0.2928 | 0.3354 | 0.2686 | 0.0302 | 0.2744 | 0.8073 | 0.0000 | 0.1568 |
| | 2 | RAN | 0.3029 | 0.3256 | 0.2865 | 0.0165 | 0.2765 | 0.7529 | 0.0000 | 0.1511 |
| | | SAL | 0.3058 | 0.3300 | 0.2823 | 0.0195 | 0.2782 | 0.7218 | 0.0256 | 0.1513 |
| | | STL | 0.2970 | 0.3366 | 0.2770 | 0.0280 | 0.2786 | 0.7519 | 0.0000 | 0.1475 |
| | 4 | RAN | 0.3074 | 0.3132 | 0.2995 | 0.0058 | 0.2800 | 0.7895 | 0.0000 | 0.1573 |
| | | SAL | 0.3107 | 0.3376 | 0.2755 | 0.0260 | 0.2831 | 0.7111 | 0.0241 | 0.1521 |
| | | STL | 0.2952 | 0.3361 | 0.2728 | 0.0290 | 0.2778 | 0.7627 | 0.0000 | 0.1525 |
| | 8 | RAN | 0.3086 | 0.3409 | 0.2618 | 0.0339 | 0.2812 | 0.8067 | 0.0000 | 0.1571 |
| | | SAL | 0.3044 | 0.3349 | 0.2666 | 0.0283 | 0.2781 | 0.7229 | 0.0385 | 0.1532 |
| | | STL | 0.2984 | 0.3407 | 0.2721 | 0.0302 | 0.2824 | 0.7965 | 0.0185 | 0.1503 |
| | 16 | RAN | 0.3144 | 0.3272 | 0.2895 | 0.0176 | 0.2860 | 0.8333 | 0.0267 | 0.1597 |
| | | SAL | 0.3128 | 0.3371 | 0.2813 | 0.0233 | 0.2860 | 0.7733 | 0.0233 | 0.1605 |
| | | STL | 0.3028 | 0.3411 | 0.2835 | 0.0271 | 0.2855 | 0.8696 | 0.0000 | 0.1584 |

Table H.22: grouped-nb-twitter-GB3-ALL-ALL-50

| GB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | RAN | 0.2437 | 0.2633 | 0.2241 | 0.0196 | 0.2035 | 0.8615 | 0.0000 | 0.1692 |
| | | SAL | 0.2499 | 0.2646 | 0.2351 | 0.0147 | 0.2094 | 0.7941 | 0.0000 | 0.1661 |
| | | STL | 0.2515 | 0.2862 | 0.2169 | 0.0347 | 0.2230 | 0.8125 | 0.0000 | 0.1641 |
| | 1 | RAN | 0.2693 | 0.2953 | 0.2433 | 0.0260 | 0.2433 | 0.7416 | 0.0000 | 0.1541 |
| | | SAL | 0.2637 | 0.3021 | 0.2254 | 0.0384 | 0.2394 | 0.8000 | 0.0000 | 0.1577 |
| | | STL | 0.2612 | 0.2873 | 0.2352 | 0.0260 | 0.2414 | 0.7368 | 0.0000 | 0.1533 |
| | 2 | RAN | 0.2660 | 0.3023 | 0.2297 | 0.0363 | 0.2407 | 0.7333 | 0.0000 | 0.1509 |
| | | SAL | 0.2724 | 0.2949 | 0.2500 | 0.0225 | 0.2439 | 0.7292 | 0.0000 | 0.1484 |
| | | STL | 0.2619 | 0.2891 | 0.2347 | 0.0272 | 0.2423 | 0.6869 | 0.0000 | 0.1450 |
| | 4 | RAN | 0.2686 | 0.2787 | 0.2584 | 0.0101 | 0.2427 | 0.6875 | 0.0000 | 0.1541 |
| | | SAL | 0.2696 | 0.3128 | 0.2263 | 0.0433 | 0.2442 | 0.7263 | 0.0000 | 0.1560 |
| | | STL | 0.2619 | 0.2881 | 0.2356 | 0.0262 | 0.2432 | 0.7368 | 0.0000 | 0.1523 |
| | 8 | RAN | 0.2650 | 0.2663 | 0.2638 | 0.0013 | 0.2401 | 0.7325 | 0.0196 | 0.1503 |
| | | SAL | 0.2685 | 0.2990 | 0.2381 | 0.0305 | 0.2439 | 0.7013 | 0.0235 | 0.1469 |
| | | STL | 0.2636 | 0.2903 | 0.2368 | 0.0268 | 0.2462 | 0.7478 | 0.0171 | 0.1470 |
| | 16 | RAN | 0.2788 | 0.2923 | 0.2653 | 0.0135 | 0.2526 | 0.7500 | 0.0000 | 0.1603 |
| | | SAL | 0.2787 | 0.3096 | 0.2479 | 0.0308 | 0.2524 | 0.7568 | 0.0000 | 0.1575 |
| | | STL | 0.2668 | 0.2948 | 0.2389 | 0.0280 | 0.2487 | 0.8116 | 0.0000 | 0.1554 |

Table H.23: grouped-nb-twitter-GB3-ALL-ALL-75

| GB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | RAN | 0.1973 | 0.1973 | 0.1973 | 0.0000 | 0.1600 | 0.7879 | 0.0000 | 0.1540 |
| | | SAL | 0.1995 | 0.1995 | 0.1995 | 0.0000 | 0.1619 | 0.7647 | 0.0000 | 0.1554 |
| | | STL | 0.1995 | 0.1995 | 0.1995 | 0.0000 | 0.1619 | 0.7647 | 0.0000 | 0.1554 |
| | 1 | RAN | 0.2148 | 0.2148 | 0.2148 | 0.0000 | 0.1910 | 0.7105 | 0.0000 | 0.1453 |
| | | SAL | 0.2154 | 0.2154 | 0.2154 | 0.0000 | 0.1910 | 0.6923 | 0.0000 | 0.1455 |
| | | STL | 0.2154 | 0.2154 | 0.2154 | 0.0000 | 0.1910 | 0.6923 | 0.0000 | 0.1455 |
| | 2 | RAN | 0.2186 | 0.2186 | 0.2186 | 0.0000 | 0.1956 | 0.6458 | 0.0000 | 0.1411 |
| | | SAL | 0.2191 | 0.2191 | 0.2191 | 0.0000 | 0.1940 | 0.6437 | 0.0000 | 0.1390 |
| | | STL | 0.2191 | 0.2191 | 0.2191 | 0.0000 | 0.1940 | 0.6437 | 0.0000 | 0.1390 |
| | 4 | RAN | 0.2190 | 0.2190 | 0.2190 | 0.0000 | 0.1947 | 0.6585 | 0.0000 | 0.1449 |
| | | SAL | 0.2200 | 0.2200 | 0.2200 | 0.0000 | 0.1958 | 0.6914 | 0.0000 | 0.1451 |
| | | STL | 0.2200 | 0.2200 | 0.2200 | 0.0000 | 0.1958 | 0.6914 | 0.0000 | 0.1451 |
| | 8 | RAN | 0.2187 | 0.2187 | 0.2187 | 0.0000 | 0.1953 | 0.6914 | 0.0000 | 0.1426 |
| | | SAL | 0.2176 | 0.2176 | 0.2176 | 0.0000 | 0.1936 | 0.7000 | 0.0000 | 0.1414 |
| | | STL | 0.2176 | 0.2176 | 0.2176 | 0.0000 | 0.1936 | 0.7000 | 0.0000 | 0.1414 |
| | 16 | RAN | 0.2233 | 0.2233 | 0.2233 | 0.0000 | 0.1989 | 0.6829 | 0.0000 | 0.1459 |
| | | SAL | 0.2216 | 0.2216 | 0.2216 | 0.0000 | 0.1969 | 0.7297 | 0.0000 | 0.1485 |
| | | STL | 0.2216 | 0.2216 | 0.2216 | 0.0000 | 0.1969 | 0.7297 | 0.0000 | 0.1485 |

Table H.24: grouped-nb-twitter-GB3-ALL-ALL-150

| OSB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 5 | 0 | RAN | 0.6437 | 0.7942 | 0.4983 | 0.0675 | 0.6224 | 0.9170 | 0.2813 | 0.1354 |
| | | SAL | 0.6529 | 0.8164 | 0.5263 | 0.0725 | 0.6323 | 0.9254 | 0.1481 | 0.1378 |
| | | STL | 0.6459 | 0.7520 | 0.5284 | 0.0612 | 0.6376 | 0.9175 | 0.2326 | 0.1137 |
| | 1 | RAN | 0.5594 | 0.8293 | 0.4149 | 0.1044 | 0.5264 | 0.9336 | 0.1370 | 0.1726 |
| | | SAL | 0.5725 | 0.7768 | 0.3913 | 0.0853 | 0.5334 | 0.9419 | 0.0597 | 0.1610 |
| | | STL | 0.5564 | 0.7425 | 0.4484 | 0.0694 | 0.5290 | 0.9064 | 0.0896 | 0.1580 |
| | 2 | RAN | 0.5741 | 0.7734 | 0.4230 | 0.0814 | 0.5321 | 0.9243 | 0.0597 | 0.1702 |
| | | SAL | 0.5746 | 0.7698 | 0.3836 | 0.0949 | 0.5340 | 0.9172 | 0.0597 | 0.1770 |
| | | STL | 0.5575 | 0.7412 | 0.4452 | 0.0705 | 0.5323 | 0.9045 | 0.0896 | 0.1544 |
| | 4 | RAN | 0.5555 | 0.7413 | 0.4019 | 0.0820 | 0.5197 | 0.9315 | 0.0800 | 0.1651 |
| | | SAL | 0.5734 | 0.7768 | 0.3836 | 0.1005 | 0.5335 | 0.9419 | 0.0597 | 0.1787 |
| | | STL | 0.5592 | 0.7439 | 0.4484 | 0.0704 | 0.5344 | 0.9064 | 0.1270 | 0.1515 |
| | 8 | RAN | 0.5613 | 0.7762 | 0.3429 | 0.0956 | 0.5209 | 0.9160 | 0.0896 | 0.1660 |
| | | SAL | 0.5662 | 0.7790 | 0.3942 | 0.1056 | 0.5261 | 0.9419 | 0.0597 | 0.1794 |
| | | STL | 0.5602 | 0.7439 | 0.4484 | 0.0692 | 0.5346 | 0.9064 | 0.1449 | 0.1566 |
| | 16 | RAN | 0.5534 | 0.7395 | 0.3857 | 0.0860 | 0.5170 | 0.9427 | 0.0351 | 0.1675 |
| | | SAL | 0.5669 | 0.7479 | 0.4126 | 0.0824 | 0.5264 | 0.9458 | 0.0606 | 0.1700 |
| | | STL | 0.5559 | 0.7425 | 0.4484 | 0.0706 | 0.5318 | 0.9064 | 0.0896 | 0.1524 |

Table H.25: grouped-nb-twitter-OSB3-ALL-ALL-5

| OSB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 10 | 0 | RAN | 0.5287 | 0.6236 | 0.4526 | 0.0468 | 0.5043 | 0.9069 | 0.0988 | 0.1461 |
| | | SAL | 0.5283 | 0.6316 | 0.4333 | 0.0606 | 0.5052 | 0.8555 | 0.1481 | 0.1516 |
| | | STL | 0.5242 | 0.6410 | 0.4355 | 0.0552 | 0.5136 | 0.8996 | 0.1075 | 0.1380 |
| | 1 | RAN | 0.4405 | 0.6071 | 0.3160 | 0.0821 | 0.4016 | 0.9122 | 0.0000 | 0.1771 |
| | | SAL | 0.4509 | 0.5714 | 0.3189 | 0.0666 | 0.4070 | 0.9018 | 0.0000 | 0.1669 |
| | | STL | 0.4247 | 0.5138 | 0.3461 | 0.0495 | 0.3844 | 0.8610 | 0.0000 | 0.1723 |
| | 2 | RAN | 0.4500 | 0.5798 | 0.2883 | 0.0749 | 0.4064 | 0.8989 | 0.0000 | 0.1790 |
| | | SAL | 0.4516 | 0.5840 | 0.3535 | 0.0755 | 0.4076 | 0.9098 | 0.0000 | 0.1744 |
| | | STL | 0.4245 | 0.5163 | 0.3461 | 0.0494 | 0.3866 | 0.8591 | 0.0000 | 0.1712 |
| | 4 | RAN | 0.4429 | 0.6238 | 0.3309 | 0.0860 | 0.3974 | 0.8846 | 0.0000 | 0.1835 |
| | | SAL | 0.4499 | 0.5940 | 0.2876 | 0.0863 | 0.4039 | 0.9228 | 0.0000 | 0.1790 |
| | | STL | 0.4281 | 0.5172 | 0.3455 | 0.0505 | 0.3899 | 0.8610 | 0.0000 | 0.1700 |
| | 8 | RAN | 0.4480 | 0.5315 | 0.3264 | 0.0581 | 0.4033 | 0.8857 | 0.0000 | 0.1802 |
| | | SAL | 0.4459 | 0.5966 | 0.2853 | 0.0911 | 0.4015 | 0.9228 | 0.0000 | 0.1778 |
| | | STL | 0.4289 | 0.5172 | 0.3461 | 0.0490 | 0.3902 | 0.8610 | 0.0000 | 0.1729 |
| | 16 | RAN | 0.4520 | 0.5954 | 0.3539 | 0.0628 | 0.4055 | 0.8785 | 0.0317 | 0.1824 |
| | | SAL | 0.4475 | 0.5826 | 0.3245 | 0.0693 | 0.4015 | 0.9265 | 0.0000 | 0.1731 |
| | | STL | 0.4278 | 0.5163 | 0.3438 | 0.0470 | 0.3902 | 0.8610 | 0.0000 | 0.1712 |

Table H.26: grouped-nb-twitter-OSB3-ALL-ALL-10

| OSB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 25 | 0 | RAN | 0.4008 | 0.4626 | 0.3364 | 0.0404 | 0.3727 | 0.8824 | 0.0000 | 0.1687 |
| | | SAL | 0.3985 | 0.4848 | 0.3207 | 0.0592 | 0.3731 | 0.8219 | 0.0000 | 0.1681 |
| | | STL | 0.3945 | 0.4509 | 0.3359 | 0.0360 | 0.3798 | 0.8529 | 0.0215 | 0.1572 |
| | 1 | RAN | 0.3270 | 0.4237 | 0.2419 | 0.0557 | 0.2795 | 0.8238 | 0.0000 | 0.1742 |
| | | SAL | 0.3207 | 0.3673 | 0.2870 | 0.0337 | 0.2706 | 0.8341 | 0.0000 | 0.1677 |
| | | STL | 0.3142 | 0.3626 | 0.2818 | 0.0254 | 0.2730 | 0.7892 | 0.0000 | 0.1646 |
| | 2 | RAN | 0.3274 | 0.3719 | 0.2897 | 0.0267 | 0.2796 | 0.8333 | 0.0000 | 0.1735 |
| | | SAL | 0.3248 | 0.3951 | 0.2704 | 0.0501 | 0.2789 | 0.8386 | 0.0000 | 0.1704 |
| | | STL | 0.3146 | 0.3618 | 0.2824 | 0.0259 | 0.2742 | 0.7768 | 0.0000 | 0.1643 |
| | 4 | RAN | 0.3257 | 0.3685 | 0.2773 | 0.0365 | 0.2736 | 0.7942 | 0.0000 | 0.1666 |
| | | SAL | 0.3251 | 0.3814 | 0.2859 | 0.0424 | 0.2764 | 0.8571 | 0.0000 | 0.1721 |
| | | STL | 0.3176 | 0.3655 | 0.2818 | 0.0266 | 0.2776 | 0.7830 | 0.0000 | 0.1650 |
| | 8 | RAN | 0.3303 | 0.3894 | 0.2570 | 0.0400 | 0.2864 | 0.8219 | 0.0000 | 0.1702 |
| | | SAL | 0.3280 | 0.3937 | 0.2782 | 0.0417 | 0.2810 | 0.8389 | 0.0000 | 0.1702 |
| | | STL | 0.3164 | 0.3622 | 0.2849 | 0.0250 | 0.2769 | 0.7795 | 0.0000 | 0.1639 |
| | 16 | RAN | 0.3269 | 0.4016 | 0.2749 | 0.0389 | 0.2800 | 0.8629 | 0.0000 | 0.1718 |
| | | SAL | 0.3246 | 0.3746 | 0.2837 | 0.0368 | 0.2782 | 0.8444 | 0.0000 | 0.1728 |
| | | STL | 0.3172 | 0.3659 | 0.2786 | 0.0281 | 0.2774 | 0.7792 | 0.0000 | 0.1647 |

Table H.27: grouped-nb-twitter-OSB3-ALL-ALL-25

| OSB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 50 | 0 | RAN | 0.3221 | 0.3322 | 0.3156 | 0.0072 | 0.2965 | 0.7792 | 0.0000 | 0.1680 |
| | | SAL | 0.3265 | 0.3739 | 0.2838 | 0.0369 | 0.3002 | 0.7895 | 0.0000 | 0.1637 |
| | | STL | 0.3212 | 0.3505 | 0.2914 | 0.0241 | 0.3046 | 0.7826 | 0.0000 | 0.1603 |
| | 1 | RAN | 0.2549 | 0.2620 | 0.2507 | 0.0051 | 0.2147 | 0.7119 | 0.0000 | 0.1596 |
| | | SAL | 0.2553 | 0.2797 | 0.2254 | 0.0225 | 0.2085 | 0.7556 | 0.0000 | 0.1540 |
| | | STL | 0.2504 | 0.2786 | 0.2361 | 0.0199 | 0.2113 | 0.7250 | 0.0000 | 0.1582 |
| | 2 | RAN | 0.2546 | 0.2797 | 0.2360 | 0.0184 | 0.2082 | 0.7712 | 0.0000 | 0.1553 |
| | | SAL | 0.2566 | 0.2674 | 0.2427 | 0.0103 | 0.2142 | 0.7350 | 0.0000 | 0.1598 |
| | | STL | 0.2502 | 0.2810 | 0.2332 | 0.0218 | 0.2118 | 0.7244 | 0.0000 | 0.1555 |
| | 4 | RAN | 0.2549 | 0.2751 | 0.2296 | 0.0189 | 0.2110 | 0.7511 | 0.0000 | 0.1598 |
| | | SAL | 0.2569 | 0.2670 | 0.2385 | 0.0130 | 0.2136 | 0.7412 | 0.0000 | 0.1620 |
| | | STL | 0.2524 | 0.2819 | 0.2343 | 0.0211 | 0.2132 | 0.7289 | 0.0000 | 0.1563 |
| | 8 | RAN | 0.2596 | 0.2831 | 0.2305 | 0.0218 | 0.2154 | 0.7585 | 0.0000 | 0.1604 |
| | | SAL | 0.2575 | 0.2741 | 0.2398 | 0.0140 | 0.2151 | 0.7511 | 0.0000 | 0.1583 |
| | | STL | 0.2512 | 0.2819 | 0.2314 | 0.0220 | 0.2128 | 0.7261 | 0.0000 | 0.1574 |
| | 16 | RAN | 0.2601 | 0.2913 | 0.2410 | 0.0223 | 0.2134 | 0.7609 | 0.0000 | 0.1602 |
| | | SAL | 0.2572 | 0.2795 | 0.2335 | 0.0188 | 0.2140 | 0.7429 | 0.0000 | 0.1620 |
| | | STL | 0.2518 | 0.2831 | 0.2300 | 0.0227 | 0.2133 | 0.7248 | 0.0000 | 0.1560 |

Table H.28: grouped-nb-twitter-OSB3-ALL-ALL-50

| OSB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 75 | 0 | RAN | 0.2854 | 0.3004 | 0.2704 | 0.0150 | 0.2607 | 0.7945 | 0.0000 | 0.1585 |
| | | SAL | 0.2912 | 0.3039 | 0.2786 | 0.0127 | 0.2656 | 0.7941 | 0.0000 | 0.1645 |
| | | STL | 0.2815 | 0.3048 | 0.2582 | 0.0233 | 0.2638 | 0.7887 | 0.0000 | 0.1585 |
| | 1 | RAN | 0.2241 | 0.2444 | 0.2038 | 0.0203 | 0.1842 | 0.7364 | 0.0000 | 0.1536 |
| | | SAL | 0.2200 | 0.2427 | 0.1972 | 0.0228 | 0.1767 | 0.7042 | 0.0000 | 0.1510 |
| | | STL | 0.2196 | 0.2347 | 0.2044 | 0.0152 | 0.1844 | 0.6905 | 0.0000 | 0.1493 |
| | 2 | RAN | 0.2285 | 0.2345 | 0.2226 | 0.0060 | 0.1854 | 0.8051 | 0.0000 | 0.1548 |
| | | SAL | 0.2230 | 0.2654 | 0.1806 | 0.0424 | 0.1796 | 0.7176 | 0.0000 | 0.1496 |
| | | STL | 0.2210 | 0.2378 | 0.2042 | 0.0168 | 0.1868 | 0.7160 | 0.0000 | 0.1493 |
| | 4 | RAN | 0.2268 | 0.2464 | 0.2072 | 0.0196 | 0.1886 | 0.7229 | 0.0000 | 0.1523 |
| | | SAL | 0.2233 | 0.2615 | 0.1851 | 0.0382 | 0.1797 | 0.7213 | 0.0000 | 0.1506 |
| | | STL | 0.2209 | 0.2383 | 0.2035 | 0.0174 | 0.1869 | 0.7073 | 0.0000 | 0.1504 |
| | 8 | RAN | 0.2212 | 0.2226 | 0.2198 | 0.0014 | 0.1824 | 0.6875 | 0.0000 | 0.1489 |
| | | SAL | 0.2240 | 0.2611 | 0.1870 | 0.0371 | 0.1812 | 0.7077 | 0.0000 | 0.1508 |
| | | STL | 0.2205 | 0.2376 | 0.2033 | 0.0172 | 0.1859 | 0.7073 | 0.0000 | 0.1502 |
| | 16 | RAN | 0.2274 | 0.2482 | 0.2065 | 0.0209 | 0.1861 | 0.7330 | 0.0000 | 0.1534 |
| | | SAL | 0.2204 | 0.2435 | 0.1973 | 0.0231 | 0.1791 | 0.7360 | 0.0000 | 0.1509 |
| | | STL | 0.2207 | 0.2382 | 0.2033 | 0.0174 | 0.1863 | 0.6905 | 0.0000 | 0.1501 |

Table H.29: grouped-nb-twitter-OSB3-ALL-ALL-75

| OSB3 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Group Type | Accuracy | | | | F-Score | | | |
| | | | AVG | MAX | MIN | STDEV | AVG | MAX | MIN | STDEV |
| 150 | 0 | RAN | 0.2320 | 0.2320 | 0.2320 | 0.0000 | 0.2087 | 0.7778 | 0.0000 | 0.1539 |
| | | SAL | 0.2337 | 0.2337 | 0.2337 | 0.0000 | 0.2113 | 0.7467 | 0.0000 | 0.1571 |
| | | STL | 0.2337 | 0.2337 | 0.2337 | 0.0000 | 0.2113 | 0.7467 | 0.0000 | 0.1571 |
| | 1 | RAN | 0.1773 | 0.1773 | 0.1773 | 0.0000 | 0.1407 | 0.6337 | 0.0000 | 0.1319 |
| | | SAL | 0.1776 | 0.1776 | 0.1776 | 0.0000 | 0.1400 | 0.6058 | 0.0000 | 0.1311 |
| | | STL | 0.1776 | 0.1776 | 0.1776 | 0.0000 | 0.1400 | 0.6058 | 0.0000 | 0.1311 |
| | 2 | RAN | 0.1792 | 0.1792 | 0.1792 | 0.0000 | 0.1413 | 0.6250 | 0.0000 | 0.1336 |
| | | SAL | 0.1790 | 0.1790 | 0.1790 | 0.0000 | 0.1427 | 0.6058 | 0.0000 | 0.1316 |
| | | STL | 0.1790 | 0.1790 | 0.1790 | 0.0000 | 0.1427 | 0.6058 | 0.0000 | 0.1316 |
| | 4 | RAN | 0.1788 | 0.1788 | 0.1788 | 0.0000 | 0.1413 | 0.6244 | 0.0000 | 0.1336 |
| | | SAL | 0.1784 | 0.1784 | 0.1784 | 0.0000 | 0.1412 | 0.6184 | 0.0000 | 0.1325 |
| | | STL | 0.1784 | 0.1784 | 0.1784 | 0.0000 | 0.1412 | 0.6184 | 0.0000 | 0.1325 |
| | 8 | RAN | 0.1810 | 0.1810 | 0.1810 | 0.0000 | 0.1442 | 0.6570 | 0.0000 | 0.1339 |
| | | SAL | 0.1787 | 0.1787 | 0.1787 | 0.0000 | 0.1412 | 0.6087 | 0.0000 | 0.1322 |
| | | STL | 0.1787 | 0.1787 | 0.1787 | 0.0000 | 0.1412 | 0.6087 | 0.0000 | 0.1322 |
| | 16 | RAN | 0.1797 | 0.1797 | 0.1797 | 0.0000 | 0.1416 | 0.6540 | 0.0000 | 0.1334 |
| | | SAL | 0.1785 | 0.1785 | 0.1785 | 0.0000 | 0.1408 | 0.6377 | 0.0000 | 0.1326 |
| | | STL | 0.1785 | 0.1785 | 0.1785 | 0.0000 | 0.1408 | 0.6377 | 0.0000 | 0.1326 |

Table H.30: grouped-nb-twitter-OSB3-ALL-ALL-150

# APPENDIX I:
# Liblinear Scores (Accuracy / Size) for the ENRON Email Corpus

| GM1 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 5 | 0 | 0.4374 | 0.8269 | 1.8907 |
| | 1 | 0.2700 | 0.8233 | 3.0491 |
| | 2 | 0.1440 | 0.8216 | 5.7048 |
| | 4 | 0.0754 | 0.8298 | 11.0081 |
| | 8 | 0.0384 | 0.8298 | 21.6131 |
| | 16 | 0.0192 | 0.8239 | 42.8058 |

Table I.1: liblinear-enron-GM1-ALL-ALL-5

| GM1 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 10 | 0 | 0.2789 | 0.7611 | 2.7294 |
| | 1 | 0.1509 | 0.7610 | 5.0430 |
| | 2 | 0.0839 | 0.7594 | 9.0498 |
| | 4 | 0.0445 | 0.7602 | 17.0733 |
| | 8 | 0.0229 | 0.7578 | 33.1029 |
| | 16 | 0.0117 | 0.7622 | 65.0962 |

Table I.2: liblinear-enron-GM1-ALL-ALL-10

| GM1 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 25 | 0 | 0.1124 | 0.6845 | 6.0891 |
| | 1 | 0.0602 | 0.6847 | 11.3773 |
| | 2 | 0.0353 | 0.6873 | 19.4874 |
| | 4 | 0.0191 | 0.6819 | 35.6598 |
| | 8 | 0.0101 | 0.6862 | 67.9353 |
| | 16 | 0.0052 | 0.6861 | 132.4731 |

Table I.3: liblinear-enron-GM1-ALL-ALL-25

| GM1 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 50 | 0 | 0.0474 | 0.6411 | 13.5303 |
| | 1 | 0.0282 | 0.6364 | 22.5920 |
| | 2 | 0.0171 | 0.6420 | 37.5149 |
| | 4 | 0.0095 | 0.6356 | 67.2513 |
| | 8 | 0.0051 | 0.6419 | 126.8646 |
| | 16 | 0.0026 | 0.6437 | 246.0215 |

Table I.4: liblinear-enron-GM1-ALL-ALL-50

| GM1 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 75 | 0 | 0.0280 | 0.6146 | 21.9684 |
| | 1 | 0.0179 | 0.6101 | 34.1738 |
| | 2 | 0.0110 | 0.6127 | 55.8602 |
| | 4 | 0.0062 | 0.6132 | 98.2967 |
| | 8 | 0.0033 | 0.6085 | 185.8852 |
| | 16 | 0.0017 | 0.6137 | 359.8159 |

Table I.5: liblinear-enron-GM1-ALL-ALL-75

| GM1 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 150 | 0 | 0.0107 | 0.6060 | 56.6487 |
| | 1 | 0.0085 | 0.5951 | 69.9670 |
| | 2 | 0.0053 | 0.5982 | 112.2744 |
| | 4 | 0.0031 | 0.6083 | 196.3090 |
| | 8 | 0.0016 | 0.5990 | 364.9695 |
| | 16 | 0.0009 | 0.5987 | 701.2420 |

Table I.6: liblinear-enron-GM1-ALL-ALL-150

| GM2 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 5 | 0 | 0.0451 | 0.8607 | 19.0724 |
| | 1 | 0.0133 | 0.8193 | 61.7992 |
| | 2 | 0.0067 | 0.8192 | 121.3905 |
| | 4 | 0.0034 | 0.8187 | 242.0133 |
| | 8 | 0.0017 | 0.8199 | 483.7247 |
| | 16 | 0.0008 | 0.8154 | 975.2060 |

Table I.7: liblinear-enron-GM2-ALL-ALL-5

| GM2 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 10 | 0 | 0.0344 | 0.8150 | 23.7182 |
| | 1 | 0.0080 | 0.7551 | 93.8882 |
| | 2 | 0.0041 | 0.7578 | 184.0112 |
| | 4 | 0.0020 | 0.7502 | 366.0065 |
| | 8 | 0.0010 | 0.7581 | 732.8332 |
| | 16 | 0.0005 | 0.7528 | 1477.4525 |

Table I.8: liblinear-enron-GM2-ALL-ALL-10

| GM2 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 25 | 0 | 0.0173 | 0.7696 | 44.4495 |
| | 1 | 0.0036 | 0.6790 | 190.3109 |
| | 2 | 0.0018 | 0.6822 | 372.1569 |
| | 4 | 0.0009 | 0.6810 | 738.3009 |

Table I.9: liblinear-enron-GM2-ALL-ALL-25

| GM2 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 50 | 0 | 0.0078 | 0.7402 | 94.3040 |
| | 1 | 0.0018 | 0.6329 | 352.2680 |
| | 2 | 0.0009 | 0.6341 | 686.3251 |

Table I.10: liblinear-enron-GM2-ALL-ALL-50

| GM2 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 75 | 0 | 0.0047 | 0.7330 | 157.4604 |
| | 1 | 0.0012 | 0.6120 | 511.3843 |
| | 2 | 0.0006 | 0.5987 | 1001.0303 |

Table I.11: liblinear-enron-GM2-ALL-ALL-75

| GM2 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 150 | 0 | 0.0018 | 0.7447 | 407.3896 |
| | 1 | 0.0006 | 0.5978 | 1001.9953 |

Table I.12: liblinear-enron-GM2-ALL-ALL-150

| GM5 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 5 | 0 | 0.0057 | 0.6881 | 120.7109 |
| | 1 | 0.0036 | 0.8117 | 224.1507 |
| | 2 | 0.0018 | 0.8118 | 448.0959 |
| | 4 | 0.0009 | 0.8130 | 895.4960 |
| | 8 | 0.0005 | 0.8070 | 1792.4125 |

Table I.13: liblinear-enron-GM5-ALL-ALL-5

| GM5 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 10 | 0 | 0.0044 | 0.6297 | 144.1015 |
| | 1 | 0.0022 | 0.7519 | 339.2591 |
| | 2 | 0.0011 | 0.7440 | 675.1472 |
| | 4 | 0.0005 | 0.7418 | 1350.7320 |

Table I.14: liblinear-enron-GM5-ALL-ALL-10

| GM5 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 25 | 0 | 0.0021 | 0.5499 | 268.0080 |
| | 1 | 0.0010 | 0.6759 | 680.7577 |
| | 2 | 0.0005 | 0.6728 | 1369.3669 |

Table I.15: liblinear-enron-GM5-ALL-ALL-25

| GM5 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 50 | 0 | 0.0009 | 0.5323 | 581.8953 |
| | 1 | 0.0005 | 0.6259 | 1263.9173 |

Table I.16: liblinear-enron-GM5-ALL-ALL-50

| GM5 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 75 | 0 | 0.0005 | 0.5211 | 986.9350 |

Table I.17: liblinear-enron-GM5-ALL-ALL-75

| GB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 5 | 0 | 0.0138 | 0.8529 | 61.9115 |
| | 1 | 0.0068 | 0.8494 | 124.5811 |
| | 2 | 0.0035 | 0.8476 | 240.5565 |
| | 4 | 0.0018 | 0.8579 | 475.1952 |
| | 8 | 0.0009 | 0.8536 | 946.3660 |
| | 16 | 0.0005 | 0.8523 | 1883.6282 |

Table I.18: liblinear-enron-GB3-ALL-ALL-5

| GB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 10 | 0 | 0.0106 | 0.8124 | 76.9171 |
| | 1 | 0.0043 | 0.8127 | 190.2863 |
| | 2 | 0.0022 | 0.8128 | 371.0244 |
| | 4 | 0.0011 | 0.8096 | 728.4382 |
| | 8 | 0.0006 | 0.8134 | 1431.6209 |

Table I.19: liblinear-enron-GB3-ALL-ALL-10

| GB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 25 | 0 | 0.0054 | 0.7680 | 143.4930 |
| | 1 | 0.0018 | 0.7652 | 415.2456 |
| | 2 | 0.0010 | 0.7684 | 770.3589 |

Table I.20: liblinear-enron-GB3-ALL-ALL-25

| GB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 50 | 0 | 0.0024 | 0.7409 | 302.6619 |
| | 1 | 0.0009 | 0.7372 | 799.7790 |

Table I.21: liblinear-enron-GB3-ALL-ALL-50

| GB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 75 | 0 | 0.0014 | 0.7300 | 507.8629 |
| | 1 | 0.0006 | 0.7336 | 1186.7074 |

Table I.22: liblinear-enron-GB3-ALL-ALL-75

| OSB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 5 | 0 | 0.0065 | 0.8690 | 133.2758 |
| | 1 | 0.0019 | 0.8667 | 459.2738 |
| | 2 | 0.0009 | 0.8645 | 920.9545 |
| | 4 | 0.0005 | 0.8687 | 1831.4047 |

Table I.23: liblinear-enron-OSB3-ALL-ALL-5

| OSB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 10 | 0 | 0.0050 | 0.8250 | 164.5371 |

Table I.24: liblinear-enron-OSB3-ALL-ALL-10

| OSB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 25 | 0 | 0.0026 | 0.7826 | 301.6652 |

Table I.25: liblinear-enron-OSB3-ALL-ALL-25

| OSB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 50 | 0 | 0.0012 | 0.7567 | 619.0632 |

Table I.26: liblinear-enron-OSB3-ALL-ALL-50

| OSB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 75 | 0 | 0.0007 | 0.7470 | 1035.0252 |

Table I.27: liblinear-enron-OSB3-ALL-ALL-75

THIS PAGE INTENTIONALLY LEFT BLANK

# APPENDIX J: Liblinear Scores (Accuracy / Size) for the Twitter Short Message Corpus

| GM1 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 5 | 0 | 2.1731 | 0.6212 | 0.2859 |
| | 1 | 0.2283 | 0.6228 | 2.7276 |
| | 2 | 0.1142 | 0.6147 | 5.3818 |
| | 4 | 0.0578 | 0.6172 | 10.6855 |
| | 8 | 0.0294 | 0.6273 | 21.3067 |
| | 16 | 0.0145 | 0.6181 | 42.4984 |

Table J.1: liblinear-twitter-GM1-ALL-ALL-5

| GM1 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 10 | 0 | 1.0850 | 0.4762 | 0.4389 |
| | 1 | 0.1143 | 0.4813 | 4.2117 |
| | 2 | 0.0589 | 0.4845 | 8.2212 |
| | 4 | 0.0298 | 0.4841 | 16.2411 |
| | 8 | 0.0149 | 0.4816 | 32.3032 |
| | 16 | 0.0075 | 0.4800 | 64.3495 |

Table J.2: liblinear-twitter-GM1-ALL-ALL-10

| GM1 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 25 | 0 | 0.3301 | 0.3461 | 1.0483 |
| | 1 | 0.0390 | 0.3408 | 8.7461 |
| | 2 | 0.0206 | 0.3465 | 16.8384 |
| | 4 | 0.0106 | 0.3510 | 33.0149 |
| | 8 | 0.0052 | 0.3419 | 65.3817 |
| | 16 | 0.0026 | 0.3411 | 130.0523 |

Table J.3: liblinear-twitter-GM1-ALL-ALL-25

| GM1 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 50 | 0 | 0.1153 | 0.2693 | 2.3366 |
| | 1 | 0.0165 | 0.2704 | 16.4318 |
| | 2 | 0.0086 | 0.2705 | 31.3058 |
| | 4 | 0.0044 | 0.2710 | 61.0727 |
| | 8 | 0.0022 | 0.2712 | 120.6237 |
| | 16 | 0.0011 | 0.2710 | 239.7010 |

Table J.4: liblinear-twitter-GM1-ALL-ALL-50

| GM1 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 75 | 0 | 0.0585 | 0.2273 | 3.8856 |
| | 1 | 0.0095 | 0.2293 | 24.1926 |
| | 2 | 0.0051 | 0.2318 | 45.8821 |
| | 4 | 0.0026 | 0.2310 | 89.2325 |
| | 8 | 0.0013 | 0.2354 | 175.9403 |
| | 16 | 0.0007 | 0.2368 | 349.3536 |

Table J.5: liblinear-twitter-GM1-ALL-ALL-75

| GM1 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 150 | 0 | 0.0194 | 0.1851 | 9.5448 |
| | 1 | 0.0040 | 0.1888 | 47.7363 |
| | 2 | 0.0020 | 0.1829 | 89.8071 |
| | 4 | 0.0011 | 0.1921 | 173.9928 |
| | 8 | 0.0006 | 0.1893 | 342.2109 |
| | 16 | 0.0003 | 0.1877 | 678.6646 |

Table J.6: liblinear-twitter-GM1-ALL-ALL-150

| GM2 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 5 | 0 | 0.3503 | 0.4844 | 1.3830 |
| | 1 | 0.0101 | 0.6241 | 61.5521 |
| | 2 | 0.0051 | 0.6221 | 121.0065 |
| | 4 | 0.0026 | 0.6253 | 241.9295 |
| | 8 | 0.0013 | 0.6282 | 483.6331 |
| | 16 | 0.0006 | 0.6258 | 976.1899 |

Table J.7: liblinear-twitter-GM2-ALL-ALL-5

| GM2 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 10 | 0 | 0.1911 | 0.3700 | 1.9363 |
| | 1 | 0.0053 | 0.4904 | 93.1498 |
| | 2 | 0.0027 | 0.4903 | 183.0523 |
| | 4 | 0.0014 | 0.4962 | 365.9331 |
| | 8 | 0.0007 | 0.4842 | 731.3270 |
| | 16 | 0.0003 | 0.4891 | 1475.9678 |

Table J.8: liblinear-twitter-GM2-ALL-ALL-10

| GM2 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 25 | 0 | 0.0600 | 0.2641 | 4.4027 |
| | 1 | 0.0018 | 0.3468 | 188.0281 |
| | 2 | 0.0009 | 0.3464 | 369.2436 |
| | 4 | 0.0005 | 0.3526 | 738.0104 |

Table J.9: liblinear-twitter-GM2-ALL-ALL-25

| GM2 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 50 | 0 | 0.0210 | 0.2097 | 9.9997 |
| | 1 | 0.0008 | 0.2679 | 346.2943 |
| | 2 | 0.0004 | 0.2707 | 679.7548 |

Table J.10: liblinear-twitter-GM2-ALL-ALL-50

| GM2 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 75 | 0 | 0.0107 | 0.1800 | 16.7491 |
| | 1 | 0.0005 | 0.2350 | 504.6377 |
| | 2 | 0.0002 | 0.2354 | 990.5135 |

Table J.11: liblinear-twitter-GM2-ALL-ALL-75

| GM2 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 150 | 0 | 0.0035 | 0.1525 | 43.1187 |
| | 1 | 0.0002 | 0.1889 | 979.9898 |

Table J.12: liblinear-twitter-GM2-ALL-ALL-150

| GM5 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 5 | 0 | 0.0814 | 0.2764 | 3.3953 |
| | 1 | 0.0026 | 0.5868 | 223.5723 |
| | 2 | 0.0013 | 0.5802 | 447.1519 |

Table J.13: liblinear-twitter-GM5-ALL-ALL-5

| GM5 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 10 | 0 | 0.0398 | 0.1718 | 4.3228 |
| | 1 | 0.0013 | 0.4383 | 338.7141 |
| | 2 | 0.0006 | 0.4382 | 677.8461 |

Table J.14: liblinear-twitter-GM5-ALL-ALL-10

| GM5 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 25 | 0 | 0.0109 | 0.1060 | 9.7684 |
| | 1 | 0.0004 | 0.3011 | 682.9173 |

Table J.15: liblinear-twitter-GM5-ALL-ALL-25

| GM5 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 50 | 0 | 0.0033 | 0.0805 | 24.4182 |

Table J.16: liblinear-twitter-GM5-ALL-ALL-50

| GM5 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 75 | 0 | 0.0015 | 0.0643 | 42.2738 |

Table J.17: liblinear-twitter-GM5-ALL-ALL-75

| GM5 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 150 | 0 | 0.0006 | 0.0636 | 114.3282 |

Table J.18: liblinear-twitter-GM5-ALL-ALL-150

| GB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 5 | 0 | 0.1364 | 0.5405 | 3.9613 |
| | 1 | 0.0045 | 0.5312 | 118.6950 |
| | 2 | 0.0023 | 0.5447 | 237.0721 |
| | 4 | 0.0011 | 0.5399 | 473.4116 |
| | 8 | 0.0006 | 0.5404 | 946.5430 |
| | 16 | 0.0003 | 0.5386 | 1893.1607 |

Table J.19: liblinear-twitter-GB3-ALL-ALL-5

| GB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 10 | 0 | 0.0737 | 0.4231 | 5.7396 |
| | 1 | 0.0023 | 0.4207 | 180.5003 |
| | 2 | 0.0012 | 0.4221 | 359.6409 |
| | 4 | 0.0006 | 0.4226 | 717.1286 |
| | 8 | 0.0003 | 0.4246 | 1431.7552 |

Table J.20: liblinear-twitter-GB3-ALL-ALL-10

| GB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 25 | 0 | 0.0228 | 0.3123 | 13.6973 |
| | 1 | 0.0008 | 0.3087 | 366.6181 |
| | 2 | 0.0004 | 0.3134 | 728.8935 |

Table J.21: liblinear-twitter-GB3-ALL-ALL-25

| GB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 50 | 0 | 0.0077 | 0.2467 | 31.9322 |
| | 1 | 0.0004 | 0.2465 | 681.2334 |

Table J.22: liblinear-twitter-GB3-ALL-ALL-50

| GB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 75 | 0 | 0.0039 | 0.2132 | 54.4683 |
| | 1 | 0.0002 | 0.2155 | 999.3930 |

Table J.23: liblinear-twitter-GB3-ALL-ALL-75

| GB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 150 | 0 | 0.0013 | 0.1739 | 135.6456 |

Table J.24: liblinear-twitter-GB3-ALL-ALL-150

| OSB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 5 | 0 | 0.0609 | 0.5430 | 8.9203 |
| | 1 | 0.0012 | 0.5391 | 458.6561 |
| | 2 | 0.0006 | 0.5427 | 916.6625 |
| | 4 | 0.0003 | 0.5391 | 1832.3842 |

Table J.25: liblinear-twitter-OSB3-ALL-ALL-5

| OSB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 10 | 0 | 0.0338 | 0.4271 | 12.6292 |
| | 1 | 0.0006 | 0.4288 | 696.0890 |
| | 2 | 0.0003 | 0.4255 | 1387.3440 |

Table J.26: liblinear-twitter-OSB3-ALL-ALL-10

| OSB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 25 | 0 | 0.0103 | 0.3084 | 29.8123 |

Table J.27: liblinear-twitter-OSB3-ALL-ALL-25

| OSB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 50 | 0 | 0.0036 | 0.2520 | 69.4942 |

Table J.28: liblinear-twitter-OSB3-ALL-ALL-50

| OSB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 75 | 0 | 0.0019 | 0.2211 | 116.9183 |

Table J.29: liblinear-twitter-OSB3-ALL-ALL-75

| OSB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 150 | 0 | 0.0006 | 0.1750 | 296.1385 |

Table J.30: liblinear-twitter-OSB3-ALL-ALL-150

# APPENDIX K:
# Naive Bayes Scores (Accuracy / Size) for the ENRON Email Corpus

| GM1 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 5 | 0 | 0.4495 | 0.7215 | 1.6050 |
| | 1 | 0.2262 | 0.6441 | 2.8472 |
| | 2 | 0.1212 | 0.6505 | 5.3662 |
| | 4 | 0.0635 | 0.6610 | 10.4023 |
| | 8 | 0.0319 | 0.6534 | 20.4755 |
| | 16 | 0.0164 | 0.6663 | 40.6210 |

Table K.1: nb-enron-GM1-ALL-ALL-5

| GM1 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 10 | 0 | 0.3186 | 0.5768 | 1.8103 |
| | 1 | 0.1634 | 0.5189 | 3.1761 |
| | 2 | 0.0915 | 0.5215 | 5.6963 |
| | 4 | 0.0504 | 0.5406 | 10.7329 |
| | 8 | 0.0257 | 0.5349 | 20.8064 |
| | 16 | 0.0131 | 0.5377 | 40.9520 |

Table K.2: nb-enron-GM1-ALL-ALL-10

| GM1 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 25 | 0 | 0.1683 | 0.4083 | 2.4263 |
| | 1 | 0.0950 | 0.3956 | 4.1628 |
| | 2 | 0.0604 | 0.4037 | 6.6867 |
| | 4 | 0.0351 | 0.4111 | 11.7248 |
| | 8 | 0.0189 | 0.4127 | 21.7989 |
| | 16 | 0.0099 | 0.4166 | 41.9447 |

Table K.3: nb-enron-GM1-ALL-ALL-25

| GM1 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 50 | 0 | 0.0903 | 0.3126 | 3.4631 |
| | 1 | 0.0543 | 0.3153 | 5.8074 |
| | 2 | 0.0383 | 0.3191 | 8.3373 |
| | 4 | 0.0248 | 0.3320 | 13.3779 |
| | 8 | 0.0141 | 0.3296 | 23.4531 |
| | 16 | 0.0078 | 0.3391 | 43.5993 |

Table K.4: nb-enron-GM1-ALL-ALL-50

| GM1 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 75 | 0 | 0.0648 | 0.2912 | 4.4965 |
| | 1 | 0.0353 | 0.2627 | 7.4519 |
| | 2 | 0.0280 | 0.2800 | 9.9879 |
| | 4 | 0.0184 | 0.2761 | 15.0311 |
| | 8 | 0.0113 | 0.2833 | 25.1073 |
| | 16 | 0.0064 | 0.2884 | 45.2539 |

Table K.5: nb-enron-GM1-ALL-ALL-75

| GM1 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 150 | 0 | 0.0323 | 0.2451 | 7.5810 |
| | 1 | 0.0149 | 0.1840 | 12.3856 |
| | 2 | 0.0127 | 0.1898 | 14.9398 |
| | 4 | 0.0098 | 0.1955 | 19.9905 |
| | 8 | 0.0066 | 0.1990 | 30.0698 |
| | 16 | 0.0040 | 0.2024 | 50.2177 |

Table K.6: nb-enron-GM1-ALL-ALL-150

| GM2 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 5 | 0 | 0.0461 | 0.8061 | 17.4811 |
| | 1 | 0.0111 | 0.6536 | 58.6657 |
| | 2 | 0.0062 | 0.7111 | 115.1229 |
| | 4 | 0.0032 | 0.7320 | 229.9337 |
| | 8 | 0.0017 | 0.7961 | 459.5158 |
| | 16 | 0.0009 | 0.8158 | 926.7927 |

Table K.7: nb-enron-GM2-ALL-ALL-5

| GM2 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 10 | 0 | 0.0393 | 0.7209 | 18.3377 |
| | 1 | 0.0092 | 0.5399 | 58.9909 |
| | 2 | 0.0051 | 0.5847 | 115.4482 |
| | 4 | 0.0027 | 0.6218 | 230.2592 |
| | 8 | 0.0016 | 0.7130 | 459.8414 |
| | 16 | 0.0008 | 0.7401 | 927.1183 |

Table K.8: nb-enron-GM2-ALL-ALL-10

| GM2 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 25 | 0 | 0.0290 | 0.6083 | 20.9885 |
| | 1 | 0.0069 | 0.4166 | 59.9664 |
| | 2 | 0.0040 | 0.4604 | 116.4243 |
| | 4 | 0.0022 | 0.5015 | 231.2357 |
| | 8 | 0.0013 | 0.6042 | 460.8180 |
| | 16 | 0.0007 | 0.6469 | 928.0949 |

Table K.9: nb-enron-GM2-ALL-ALL-25

| GM2 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 50 | 0 | 0.0214 | 0.5414 | 25.3013 |
| | 1 | 0.0056 | 0.3448 | 61.5922 |
| | 2 | 0.0032 | 0.3831 | 118.0511 |
| | 4 | 0.0018 | 0.4259 | 232.8632 |
| | 8 | 0.0012 | 0.5386 | 462.4457 |
| | 16 | 0.0006 | 0.5891 | 929.7227 |

Table K.10: nb-enron-GM2-ALL-ALL-50

| GM2 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 75 | 0 | 0.0170 | 0.5056 | 29.8167 |
| | 1 | 0.0046 | 0.2896 | 63.2180 |
| | 2 | 0.0027 | 0.3286 | 119.6779 |
| | 4 | 0.0016 | 0.3762 | 234.4907 |
| | 8 | 0.0011 | 0.5018 | 464.0733 |
| | 16 | 0.0006 | 0.5547 | 931.3504 |

Table K.11: nb-enron-GM2-ALL-ALL-75

| GM2 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 150 | 0 | 0.0106 | 0.4536 | 42.5965 |
| | 1 | 0.0032 | 0.2164 | 68.0954 |
| | 2 | 0.0021 | 0.2598 | 124.5583 |
| | 4 | 0.0013 | 0.3096 | 239.3732 |
| | 8 | 0.0010 | 0.4547 | 468.9564 |
| | 16 | 0.0005 | 0.5061 | 936.2337 |

Table K.12: nb-enron-GM2-ALL-ALL-150

| GM5 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 5 | 0 | 0.0065 | 0.7379 | 112.9389 |
| | 1 | 0.0037 | 0.7817 | 212.9081 |
| | 2 | 0.0019 | 0.8104 | 425.6461 |
| | 4 | 0.0010 | 0.8206 | 851.1609 |
| | 8 | 0.0005 | 0.8064 | 1704.9937 |
| | 16 | 0.0002 | 0.7980 | 3410.7224 |

Table K.13: nb-enron-GM5-ALL-ALL-5

| GM5 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 10 | 0 | 0.0059 | 0.6803 | 114.5784 |
| | 1 | 0.0032 | 0.6890 | 213.2250 |
| | 2 | 0.0017 | 0.7274 | 425.9630 |
| | 4 | 0.0009 | 0.7367 | 851.4779 |
| | 8 | 0.0004 | 0.7169 | 1705.3107 |
| | 16 | 0.0002 | 0.6991 | 3411.0393 |

Table K.14: nb-enron-GM5-ALL-ALL-10

| GM5 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 25 | 0 | 0.0051 | 0.6081 | 119.3005 |
| | 1 | 0.0027 | 0.5847 | 214.1756 |
| | 2 | 0.0015 | 0.6358 | 426.9138 |
| | 4 | 0.0008 | 0.6445 | 852.4287 |
| | 8 | 0.0004 | 0.5994 | 1706.2616 |
| | 16 | 0.0002 | 0.5644 | 3411.9902 |

Table K.15: nb-enron-GM5-ALL-ALL-25

| GM5 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 50 | 0 | 0.0045 | 0.5742 | 127.4399 |
| | 1 | 0.0024 | 0.5103 | 215.7599 |
| | 2 | 0.0013 | 0.5726 | 428.4984 |
| | 4 | 0.0007 | 0.5775 | 854.0134 |
| | 8 | 0.0003 | 0.5142 | 1707.8464 |
| | 16 | 0.0001 | 0.4650 | 3413.5750 |

Table K.16: nb-enron-GM5-ALL-ALL-50

| GM5 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 75 | 0 | 0.0042 | 0.5646 | 135.6983 |
| | 1 | 0.0022 | 0.4722 | 217.3442 |
| | 2 | 0.0013 | 0.5391 | 430.0830 |
| | 4 | 0.0006 | 0.5539 | 855.5982 |
| | 8 | 0.0003 | 0.4791 | 1709.4312 |
| | 16 | 0.0001 | 0.4316 | 3415.1599 |

Table K.17: nb-enron-GM5-ALL-ALL-75

| GM5 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 150 | 0 | 0.0035 | 0.5659 | 160.2268 |
| | 1 | 0.0019 | 0.4139 | 222.0971 |
| | 2 | 0.0011 | 0.4941 | 434.8369 |
| | 4 | 0.0006 | 0.5126 | 860.3524 |
| | 8 | 0.0003 | 0.4287 | 1714.1856 |
| | 16 | 0.0001 | 0.2613 | 3419.9143 |

Table K.18: nb-enron-GM5-ALL-ALL-150

| GB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 5 | 0 | 0.0138 | 0.7882 | 56.9801 |
| | 1 | 0.0070 | 0.8167 | 116.4139 |
| | 2 | 0.0036 | 0.8314 | 228.6381 |
| | 4 | 0.0019 | 0.8629 | 453.0841 |
| | 8 | 0.0010 | 0.8601 | 901.9044 |
| | 16 | 0.0005 | 0.8589 | 1800.2036 |

Table K.19: nb-enron-GB3-ALL-ALL-5

| GB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 10 | 0 | 0.0118 | 0.7057 | 59.6121 |
| | 1 | 0.0063 | 0.7586 | 120.6296 |
| | 2 | 0.0033 | 0.7729 | 232.8820 |
| | 4 | 0.0018 | 0.8070 | 457.3419 |
| | 8 | 0.0009 | 0.8074 | 906.1696 |
| | 16 | 0.0004 | 0.8091 | 1804.4721 |

Table K.20: nb-enron-GB3-ALL-ALL-10

| GB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 25 | 0 | 0.0089 | 0.5999 | 67.7286 |
| | 1 | 0.0052 | 0.6887 | 133.2767 |
| | 2 | 0.0029 | 0.7102 | 245.6138 |
| | 4 | 0.0016 | 0.7520 | 470.1153 |
| | 8 | 0.0008 | 0.7436 | 918.9651 |
| | 16 | 0.0004 | 0.7557 | 1817.2774 |

Table K.21: nb-enron-GB3-ALL-ALL-25

| GB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 50 | 0 | 0.0062 | 0.5059 | 80.9739 |
| | 1 | 0.0041 | 0.6391 | 154.3553 |
| | 2 | 0.0025 | 0.6740 | 266.8334 |
| | 4 | 0.0014 | 0.7097 | 491.4044 |
| | 8 | 0.0008 | 0.7083 | 940.2910 |
| | 16 | 0.0004 | 0.7185 | 1838.6197 |

Table K.22: nb-enron-GB3-ALL-ALL-50

| GB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 75 | 0 | 0.0050 | 0.4721 | 94.5262 |
| | 1 | 0.0035 | 0.6188 | 175.4338 |
| | 2 | 0.0023 | 0.6573 | 288.0529 |
| | 4 | 0.0014 | 0.6932 | 512.6934 |
| | 8 | 0.0007 | 0.6952 | 961.6169 |
| | 16 | 0.0004 | 0.7075 | 1859.9620 |

Table K.23: nb-enron-GB3-ALL-ALL-75

| GB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 150 | 0 | 0.0032 | 0.4282 | 133.7770 |
| | 1 | 0.0025 | 0.5976 | 238.6693 |
| | 2 | 0.0018 | 0.6499 | 351.7116 |
| | 4 | 0.0012 | 0.6860 | 576.5604 |
| | 8 | 0.0007 | 0.6889 | 1025.5946 |
| | 16 | 0.0004 | 0.7056 | 1923.9889 |

Table K.24: nb-enron-GB3-ALL-ALL-150

| OSB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 5 | 0 | 0.0069 | 0.8527 | 123.3505 |
| | 1 | 0.0020 | 0.8648 | 442.9132 |
| | 2 | 0.0010 | 0.8642 | 877.3292 |
| | 4 | 0.0005 | 0.8678 | 1746.3302 |
| | 8 | 0.0002 | 0.8638 | 3484.0829 |
| | 16 | 0.0001 | 0.8653 | 6966.6684 |

Table K.25: nb-enron-OSB3-ALL-ALL-5

| OSB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 10 | 0 | 0.0062 | 0.7967 | 128.6443 |
| | 1 | 0.0018 | 0.8161 | 451.4412 |
| | 2 | 0.0009 | 0.8180 | 885.8726 |
| | 4 | 0.0005 | 0.8177 | 1754.8804 |
| | 8 | 0.0002 | 0.8195 | 3492.6359 |
| | 16 | 0.0001 | 0.8186 | 6970.0526 |

Table K.26: nb-enron-OSB3-ALL-ALL-10

| OSB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 25 | 0 | 0.0050 | 0.7263 | 144.8636 |
| | 1 | 0.0016 | 0.7586 | 477.0252 |
| | 2 | 0.0008 | 0.7635 | 911.5027 |
| | 4 | 0.0004 | 0.7621 | 1780.5309 |
| | 8 | 0.0002 | 0.7613 | 3518.2948 |

Table K.27: nb-enron-OSB3-ALL-ALL-25

| OSB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 50 | 0 | 0.0040 | 0.6818 | 171.3725 |
| | 1 | 0.0009 | 0.4880 | 519.6651 |
| | 2 | 0.0004 | 0.4123 | 954.2195 |
| | 4 | 0.0004 | 0.7216 | 1823.2817 |

Table K.28: nb-enron-OSB3-ALL-ALL-50

| OSB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 75 | 0 | 0.0034 | 0.6713 | 198.4913 |
| | 1 | 0.0013 | 0.7074 | 562.3051 |
| | 2 | 0.0007 | 0.7112 | 996.9363 |
| | 4 | 0.0004 | 0.7089 | 1866.0326 |
| | 8 | 0.0002 | 0.7079 | 3603.8247 |

Table K.29: nb-enron-OSB3-ALL-ALL-75

| OSB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 150 | 0 | 0.0024 | 0.6572 | 276.9361 |
| | 1 | 0.0010 | 0.7041 | 690.2250 |
| | 2 | 0.0006 | 0.7091 | 1125.0867 |
| | 4 | 0.0004 | 0.7068 | 1994.2851 |
| | 8 | 0.0002 | 0.7101 | 3732.1196 |

Table K.30: nb-enron-OSB3-ALL-ALL-150

# APPENDIX L:
# Naive Bayes Scores (Accuracy / Size) for the Twitter Short Message Corpus

| GM1 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 5 | 0 | 2.8233 | 0.6264 | 0.2219 |
| | 1 | 0.1945 | 0.4974 | 2.5578 |
| | 2 | 0.1011 | 0.5130 | 5.0756 |
| | 4 | 0.0508 | 0.5132 | 10.1112 |
| | 8 | 0.0260 | 0.5238 | 20.1842 |
| | 16 | 0.0127 | 0.5102 | 40.3296 |

Table L.1: nb-twitter-GM1-ALL-ALL-5

| GM1 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 10 | 0 | 1.9815 | 0.4869 | 0.2457 |
| | 1 | 0.1344 | 0.3490 | 2.5972 |
| | 2 | 0.0718 | 0.3671 | 5.1151 |
| | 4 | 0.0364 | 0.3693 | 10.1508 |
| | 8 | 0.0186 | 0.3755 | 20.2239 |
| | 16 | 0.0090 | 0.3642 | 40.3693 |

Table L.2: nb-twitter-GM1-ALL-ALL-10

| GM1 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 25 | 0 | 1.0593 | 0.3357 | 0.3169 |
| | 1 | 0.0832 | 0.2260 | 2.7157 |
| | 2 | 0.0439 | 0.2299 | 5.2329 |
| | 4 | 0.0227 | 0.2332 | 10.2692 |
| | 8 | 0.0118 | 0.2398 | 20.3426 |
| | 16 | 0.0057 | 0.2304 | 40.4876 |

Table L.3: nb-twitter-GM1-ALL-ALL-25

| GM1 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 50 | 0 | 0.5347 | 0.2326 | 0.4350 |
| | 1 | 0.0547 | 0.1595 | 2.9149 |
| | 2 | 0.0300 | 0.1631 | 5.4320 |
| | 4 | 0.0158 | 0.1653 | 10.4665 |
| | 8 | 0.0081 | 0.1671 | 20.5404 |
| | 16 | 0.0040 | 0.1638 | 40.6876 |

Table L.4: nb-twitter-GM1-ALL-ALL-50

| GM1 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 75 | 0 | 0.3291 | 0.1820 | 0.5530 |
| | 1 | 0.0426 | 0.1328 | 3.1132 |
| | 2 | 0.0242 | 0.1364 | 5.6325 |
| | 4 | 0.0128 | 0.1370 | 10.6678 |
| | 8 | 0.0067 | 0.1390 | 20.7389 |
| | 16 | 0.0033 | 0.1360 | 40.8821 |

Table L.5: nb-twitter-GM1-ALL-ALL-75

| GM1 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 150 | 0 | 0.1375 | 0.1252 | 0.9104 |
| | 1 | 0.0265 | 0.0982 | 3.7007 |
| | 2 | 0.0157 | 0.0979 | 6.2278 |
| | 4 | 0.0087 | 0.0975 | 11.2574 |
| | 8 | 0.0046 | 0.0986 | 21.3302 |
| | 16 | 0.0024 | 0.0983 | 41.4726 |

Table L.6: nb-twitter-GM1-ALL-ALL-150

| GM2 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 5 | 0 | 0.4866 | 0.5711 | 1.1738 |
| | 1 | 0.0090 | 0.5253 | 58.3788 |
| | 2 | 0.0046 | 0.5301 | 114.8358 |
| | 4 | 0.0024 | 0.5401 | 229.6465 |
| | 8 | 0.0012 | 0.5459 | 459.2286 |
| | 16 | 0.0006 | 0.5409 | 926.5055 |

Table L.7: nb-twitter-GM2-ALL-ALL-5

| GM2 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 10 | 0 | 0.3644 | 0.4439 | 1.2181 |
| | 1 | 0.0066 | 0.3847 | 58.4174 |
| | 2 | 0.0034 | 0.3914 | 114.8738 |
| | 4 | 0.0017 | 0.4018 | 229.6847 |
| | 8 | 0.0009 | 0.4046 | 459.2667 |
| | 16 | 0.0004 | 0.4031 | 926.5439 |

Table L.8: nb-twitter-GM2-ALL-ALL-10

| GM2 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 25 | 0 | 0.2380 | 0.3215 | 1.3508 |
| | 1 | 0.0043 | 0.2508 | 58.5325 |
| | 2 | 0.0022 | 0.2510 | 114.9893 |
| | 4 | 0.0012 | 0.2686 | 229.7997 |
| | 8 | 0.0006 | 0.2768 | 459.3819 |
| | 16 | 0.0003 | 0.2703 | 926.6588 |

Table L.9: nb-twitter-GM2-ALL-ALL-25

| GM2 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 50 | 0 | 0.1598 | 0.2509 | 1.5700 |
| | 1 | 0.0032 | 0.1854 | 58.7240 |
| | 2 | 0.0016 | 0.1843 | 115.1800 |
| | 4 | 0.0009 | 0.1980 | 229.9909 |
| | 8 | 0.0004 | 0.2058 | 459.5747 |
| | 16 | 0.0002 | 0.2023 | 926.8513 |

Table L.10: nb-twitter-GM2-ALL-ALL-50

| GM2 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 75 | 0 | 0.1207 | 0.2162 | 1.7905 |
| | 1 | 0.0026 | 0.1518 | 58.9138 |
| | 2 | 0.0013 | 0.1535 | 115.3732 |
| | 4 | 0.0007 | 0.1666 | 230.1851 |
| | 8 | 0.0004 | 0.1770 | 459.7666 |
| | 16 | 0.0002 | 0.1700 | 927.0460 |

Table L.11: nb-twitter-GM2-ALL-ALL-75

| GM2 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 150 | 0 | 0.0696 | 0.1709 | 2.4569 |
| | 1 | 0.0019 | 0.1155 | 59.4939 |
| | 2 | 0.0010 | 0.1173 | 115.9487 |
| | 4 | 0.0006 | 0.1282 | 230.7586 |
| | 8 | 0.0003 | 0.1331 | 460.3425 |
| | 16 | 0.0001 | 0.1290 | 927.6131 |

Table L.12: nb-twitter-GM2-ALL-ALL-150

| GM5 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 5 | 0 | 0.1104 | 0.3453 | 3.1278 |
| | 1 | 0.0026 | 0.5530 | 212.6251 |
| | 2 | 0.0013 | 0.5523 | 425.3632 |
| | 4 | 0.0006 | 0.5516 | 850.8778 |
| | 8 | 0.0003 | 0.5550 | 1704.7107 |
| | 16 | 0.0002 | 0.5579 | 3410.4392 |

Table L.13: nb-twitter-GM5-ALL-ALL-5

| GM5 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 10 | 0 | 0.0759 | 0.2408 | 3.1727 |
| | 1 | 0.0020 | 0.4222 | 212.6591 |
| | 2 | 0.0010 | 0.4187 | 425.3972 |
| | 4 | 0.0005 | 0.4191 | 850.9119 |
| | 8 | 0.0002 | 0.4209 | 1704.7446 |
| | 16 | 0.0001 | 0.4319 | 3410.4732 |

Table L.14: nb-twitter-GM5-ALL-ALL-10

| GM5 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 25 | 0 | 0.0467 | 0.1543 | 3.3068 |
| | 1 | 0.0014 | 0.2990 | 212.7613 |
| | 2 | 0.0007 | 0.2923 | 425.4981 |
| | 4 | 0.0003 | 0.2935 | 851.0135 |
| | 8 | 0.0002 | 0.2929 | 1704.8464 |
| | 16 | 0.0001 | 0.2973 | 3410.5747 |

Table L.15: nb-twitter-GM5-ALL-ALL-25

| GM5 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 50 | 0 | 0.0347 | 0.1226 | 3.5298 |
| | 1 | 0.0011 | 0.2277 | 212.9303 |
| | 2 | 0.0005 | 0.2252 | 425.6690 |
| | 4 | 0.0003 | 0.2216 | 851.1809 |
| | 8 | 0.0001 | 0.2217 | 1705.0156 |
| | 16 | 0.0001 | 0.2245 | 3410.7461 |

Table L.16: nb-twitter-GM5-ALL-ALL-50

| GM5 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 75 | 0 | 0.0260 | 0.0977 | 3.7547 |
| | 1 | 0.0009 | 0.1960 | 213.0992 |
| | 2 | 0.0005 | 0.1952 | 425.8394 |
| | 4 | 0.0002 | 0.1878 | 851.3518 |
| | 8 | 0.0001 | 0.1902 | 1705.1849 |
| | 16 | 0.0001 | 0.1908 | 3410.9148 |

Table L.17: nb-twitter-GM5-ALL-ALL-75

| GM5 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 150 | 0 | 0.0178 | 0.0787 | 4.4242 |
| | 1 | 0.0007 | 0.1535 | 213.6106 |
| | 2 | 0.0004 | 0.1524 | 426.3482 |
| | 4 | 0.0002 | 0.1404 | 851.8602 |
| | 8 | 0.0001 | 0.1401 | 1705.6951 |
| | 16 | 0.0000 | 0.1424 | 3411.4199 |

Table L.18: nb-twitter-GM5-ALL-ALL-150

| GB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 5 | 0 | 0.1868 | 0.6179 | 3.3084 |
| | 1 | 0.0054 | 0.6109 | 112.4031 |
| | 2 | 0.0027 | 0.6160 | 224.5981 |
| | 4 | 0.0014 | 0.6199 | 449.0306 |
| | 8 | 0.0007 | 0.6187 | 897.8441 |
| | 16 | 0.0003 | 0.6265 | 1796.1397 |

Table L.19: nb-twitter-GB3-ALL-ALL-5

| GB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 10 | 0 | 0.1440 | 0.4948 | 3.4368 |
| | 1 | 0.0044 | 0.4975 | 112.6054 |
| | 2 | 0.0022 | 0.5015 | 224.8053 |
| | 4 | 0.0011 | 0.5011 | 449.2383 |
| | 8 | 0.0006 | 0.4999 | 898.0512 |
| | 16 | 0.0003 | 0.5041 | 1796.3457 |

Table L.20: nb-twitter-GB3-ALL-ALL-10

| GB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 25 | 0 | 0.0937 | 0.3584 | 3.8242 |
| | 1 | 0.0033 | 0.3724 | 113.2226 |
| | 2 | 0.0017 | 0.3760 | 225.4217 |
| | 4 | 0.0008 | 0.3773 | 449.8555 |
| | 8 | 0.0004 | 0.3733 | 898.6658 |
| | 16 | 0.0002 | 0.3838 | 1796.9620 |

Table L.21: nb-twitter-GB3-ALL-ALL-25

| GB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 50 | 0 | 0.0636 | 0.2839 | 4.4620 |
| | 1 | 0.0026 | 0.2984 | 114.2444 |
| | 2 | 0.0013 | 0.3019 | 226.4522 |
| | 4 | 0.0007 | 0.3044 | 450.8849 |
| | 8 | 0.0003 | 0.3038 | 899.6969 |
| | 16 | 0.0002 | 0.3100 | 1797.9942 |

Table L.22: nb-twitter-GB3-ALL-ALL-50

| GB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 75 | 0 | 0.0487 | 0.2484 | 5.1018 |
| | 1 | 0.0023 | 0.2648 | 115.2803 |
| | 2 | 0.0012 | 0.2668 | 227.4728 |
| | 4 | 0.0006 | 0.2667 | 451.9111 |
| | 8 | 0.0003 | 0.2657 | 900.7211 |
| | 16 | 0.0002 | 0.2748 | 1799.0202 |

Table L.23: nb-twitter-GB3-ALL-ALL-75

| GB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 150 | 0 | 0.0283 | 0.1988 | 7.0186 |
| | 1 | 0.0018 | 0.2152 | 118.3712 |
| | 2 | 0.0009 | 0.2190 | 230.5749 |
| | 4 | 0.0005 | 0.2197 | 455.0150 |
| | 8 | 0.0002 | 0.2180 | 903.8222 |
| | 16 | 0.0001 | 0.2221 | 1802.1037 |

Table L.24: nb-twitter-GB3-ALL-ALL-150

| OSB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 5 | 0 | 0.0856 | 0.6475 | 7.5640 |
| | 1 | 0.0013 | 0.5628 | 434.7978 |
| | 2 | 0.0007 | 0.5687 | 869.2005 |
| | 4 | 0.0003 | 0.5627 | 1738.1940 |
| | 8 | 0.0002 | 0.5626 | 3475.9425 |
| | 16 | 0.0001 | 0.5587 | 6958.5295 |

Table L.25: nb-twitter-OSB3-ALL-ALL-5

| OSB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 10 | 0 | 0.0673 | 0.5271 | 7.8277 |
| | 1 | 0.0010 | 0.4387 | 435.2158 |
| | 2 | 0.0005 | 0.4420 | 869.6175 |
| | 4 | 0.0003 | 0.4403 | 1738.6111 |
| | 8 | 0.0001 | 0.4410 | 3476.3611 |
| | 16 | 0.0001 | 0.4425 | 6958.9458 |

Table L.26: nb-twitter-OSB3-ALL-ALL-10

| OSB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 25 | 0 | 0.0462 | 0.3979 | 8.6114 |
| | 1 | 0.0007 | 0.3206 | 436.4719 |
| | 2 | 0.0004 | 0.3223 | 870.8699 |
| | 4 | 0.0002 | 0.3228 | 1739.8607 |
| | 8 | 0.0001 | 0.3249 | 3477.6101 |
| | 16 | 0.0000 | 0.3229 | 6960.2009 |

Table L.27: nb-twitter-OSB3-ALL-ALL-25

| OSB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 50 | 0 | 0.0326 | 0.3233 | 9.9031 |
| | 1 | 0.0006 | 0.2535 | 438.5595 |
| | 2 | 0.0003 | 0.2538 | 872.9681 |
| | 4 | 0.0001 | 0.2548 | 1741.9419 |
| | 8 | 0.0001 | 0.2561 | 3479.6877 |
| | 16 | 0.0000 | 0.2564 | 6962.2798 |

Table L.28: nb-twitter-OSB3-ALL-ALL-50

| OSB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 75 | 0 | 0.0255 | 0.2861 | 11.1995 |
| | 1 | 0.0005 | 0.2212 | 440.6614 |
| | 2 | 0.0003 | 0.2242 | 875.0073 |
| | 4 | 0.0001 | 0.2237 | 1744.0464 |
| | 8 | 0.0001 | 0.2219 | 3481.7927 |
| | 16 | 0.0000 | 0.2228 | 6964.3666 |

Table L.29: nb-twitter-OSB3-ALL-ALL-75

| OSB3 | | | | |
|---|---|---|---|---|
| Group Size | Web1T % | Score | Accuracy | Size(MB) |
| 150 | 0 | 0.0154 | 0.2332 | 15.0982 |
| | 1 | 0.0004 | 0.1775 | 446.8905 |
| | 2 | 0.0002 | 0.1791 | 881.2573 |
| | 4 | 0.0001 | 0.1786 | 1750.2371 |
| | 8 | 0.0001 | 0.1795 | 3488.0249 |
| | 16 | 0.0000 | 0.1789 | 6970.5827 |

Table L.30: nb-twitter-OSB3-ALL-ALL-150

# APPENDIX M:
# Liblinear Storage Requirements for the ENRON Email Corpus

| GM1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Size (MB) | | | | | | |
| | | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 5 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 1.40 | 0.49 | 1.89 |
| | 1 | 0.07 | 0.00 | 0.00 | 1.09 | 0.00 | 1.89 | 3.05 |
| | 2 | 0.14 | 0.00 | 0.00 | 2.17 | 0.00 | 3.39 | 5.70 |
| | 4 | 0.29 | 0.00 | 0.00 | 4.35 | 0.00 | 6.37 | 11.01 |
| | 8 | 0.58 | 0.00 | 0.00 | 8.70 | 0.00 | 12.34 | 21.61 |
| | 16 | 1.15 | 0.00 | 0.00 | 17.40 | 0.00 | 24.26 | 42.81 |

Table M.1: liblinear-enron-GM1-ALL-ALL-5

| GM1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Size (MB) | | | | | | |
| | | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 10 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 1.40 | 1.33 | 2.73 |
| | 1 | 0.07 | 0.00 | 0.00 | 1.09 | 0.00 | 3.88 | 5.04 |
| | 2 | 0.14 | 0.00 | 0.00 | 2.17 | 0.00 | 6.73 | 9.05 |
| | 4 | 0.29 | 0.00 | 0.00 | 4.35 | 0.00 | 12.44 | 17.07 |
| | 8 | 0.58 | 0.00 | 0.00 | 8.70 | 0.00 | 23.83 | 33.10 |
| | 16 | 1.15 | 0.00 | 0.00 | 17.40 | 0.00 | 46.55 | 65.10 |

Table M.2: liblinear-enron-GM1-ALL-ALL-10

| GM1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 25 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 1.40 | 4.69 | 6.09 |
| | 1 | 0.07 | 0.00 | 0.00 | 1.09 | 0.00 | 10.22 | 11.38 |
| | 2 | 0.14 | 0.00 | 0.00 | 2.17 | 0.00 | 17.17 | 19.49 |
| | 4 | 0.29 | 0.00 | 0.00 | 4.35 | 0.00 | 31.02 | 35.66 |
| | 8 | 0.58 | 0.00 | 0.00 | 8.70 | 0.00 | 58.66 | 67.94 |
| | 16 | 1.15 | 0.00 | 0.00 | 17.40 | 0.00 | 113.93 | 132.47 |

Table M.3: liblinear-enron-GM1-ALL-ALL-25

| GM1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 50 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 1.40 | 12.13 | 13.53 |
| | 1 | 0.07 | 0.00 | 0.00 | 1.09 | 0.00 | 21.43 | 22.59 |
| | 2 | 0.14 | 0.00 | 0.00 | 2.17 | 0.00 | 35.20 | 37.51 |
| | 4 | 0.29 | 0.00 | 0.00 | 4.35 | 0.00 | 62.62 | 67.25 |
| | 8 | 0.58 | 0.00 | 0.00 | 8.70 | 0.00 | 117.59 | 126.86 |
| | 16 | 1.15 | 0.00 | 0.00 | 17.40 | 0.00 | 227.48 | 246.02 |

Table M.4: liblinear-enron-GM1-ALL-ALL-50

| GM1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Size (MB) | | | | | | |
| | | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 75 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 1.40 | 20.57 | 21.97 |
| | 1 | 0.07 | 0.00 | 0.00 | 1.09 | 0.00 | 33.01 | 34.17 |
| | 2 | 0.14 | 0.00 | 0.00 | 2.17 | 0.00 | 53.54 | 55.86 |
| | 4 | 0.29 | 0.00 | 0.00 | 4.35 | 0.00 | 93.66 | 98.30 |
| | 8 | 0.58 | 0.00 | 0.00 | 8.70 | 0.00 | 176.61 | 185.89 |
| | 16 | 1.15 | 0.00 | 0.00 | 17.40 | 0.00 | 341.27 | 359.82 |

Table M.5: liblinear-enron-GM1-ALL-ALL-75

| GM1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Size (MB) | | | | | | |
| | | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 150 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 1.40 | 55.25 | 56.65 |
| | 1 | 0.07 | 0.00 | 0.00 | 1.09 | 0.00 | 68.81 | 69.97 |
| | 2 | 0.14 | 0.00 | 0.00 | 2.17 | 0.00 | 109.96 | 112.27 |
| | 4 | 0.29 | 0.00 | 0.00 | 4.35 | 0.00 | 191.67 | 196.31 |
| | 8 | 0.58 | 0.00 | 0.00 | 8.70 | 0.00 | 355.70 | 364.97 |
| | 16 | 1.15 | 0.00 | 0.00 | 17.40 | 0.00 | 682.70 | 701.24 |

Table M.6: liblinear-enron-GM1-ALL-ALL-150

| GM2 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 5 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 16.61 | 2.46 | 19.07 |
| | 1 | 1.67 | 0.00 | 0.00 | 25.19 | 0.00 | 34.94 | 61.80 |
| | 2 | 3.28 | 0.00 | 0.00 | 49.56 | 0.00 | 68.55 | 121.39 |
| | 4 | 6.56 | 0.00 | 0.00 | 99.13 | 0.00 | 136.32 | 242.01 |
| | 8 | 13.12 | 0.00 | 0.00 | 198.25 | 0.00 | 272.35 | 483.72 |
| | 16 | 26.47 | 0.00 | 0.00 | 400.00 | 0.00 | 548.74 | 975.21 |

Table M.7: liblinear-enron-GM2-ALL-ALL-5

| GM2 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 10 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 16.61 | 7.11 | 23.72 |
| | 1 | 1.67 | 0.00 | 0.00 | 25.19 | 0.00 | 67.03 | 93.89 |
| | 2 | 3.28 | 0.00 | 0.00 | 49.56 | 0.00 | 131.17 | 184.01 |
| | 4 | 6.56 | 0.00 | 0.00 | 99.13 | 0.00 | 260.31 | 366.01 |
| | 8 | 13.12 | 0.00 | 0.00 | 198.25 | 0.00 | 521.46 | 732.83 |
| | 16 | 26.47 | 0.00 | 0.00 | 400.00 | 0.00 | 1050.99 | 1477.45 |

Table M.8: liblinear-enron-GM2-ALL-ALL-10

| GM2 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 25 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 16.61 | 27.84 | 44.45 |
| | 1 | 1.67 | 0.00 | 0.00 | 25.19 | 0.00 | 163.46 | 190.31 |
| | 2 | 3.28 | 0.00 | 0.00 | 49.56 | 0.00 | 319.31 | 372.16 |
| | 4 | 6.56 | 0.00 | 0.00 | 99.13 | 0.00 | 632.61 | 738.30 |

Table M.9: liblinear-enron-GM2-ALL-ALL-25

| GM2 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 50 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 16.61 | 77.69 | 94.30 |
| | 1 | 1.67 | 0.00 | 0.00 | 25.19 | 0.00 | 325.41 | 352.27 |
| | 2 | 3.28 | 0.00 | 0.00 | 49.56 | 0.00 | 633.48 | 686.33 |

Table M.10: liblinear-enron-GM2-ALL-ALL-50

| GM2 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 75 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 16.61 | 140.85 | 157.46 |
| | 1 | 1.67 | 0.00 | 0.00 | 25.19 | 0.00 | 484.53 | 511.38 |
| | 2 | 3.28 | 0.00 | 0.00 | 49.56 | 0.00 | 948.19 | 1001.03 |

Table M.11: liblinear-enron-GM2-ALL-ALL-75

| GM2 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Size (MB) | | | | | | |
| | | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 150 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 16.61 | 390.78 | 407.39 |
| | 1 | 1.67 | 0.00 | 0.00 | 25.19 | 0.00 | 975.14 | 1002.00 |

Table M.12: liblinear-enron-GM2-ALL-ALL-150

| GM5 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Size (MB) | | | | | | |
| | | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 5 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 111.31 | 9.40 | 120.71 |
| | 1 | 6.07 | 0.00 | 0.00 | 91.79 | 0.00 | 126.29 | 224.15 |
| | 2 | 12.15 | 0.00 | 0.00 | 183.63 | 0.00 | 252.31 | 448.10 |
| | 4 | 24.31 | 0.00 | 0.00 | 367.35 | 0.00 | 503.84 | 895.50 |
| | 8 | 48.70 | 0.00 | 0.00 | 735.99 | 0.00 | 1007.72 | 1792.41 |

Table M.13: liblinear-enron-GM5-ALL-ALL-5

| GM5 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 10 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 111.31 | 32.79 | 144.10 |
| | 1 | 6.07 | 0.00 | 0.00 | 91.79 | 0.00 | 241.40 | 339.26 |
| | 2 | 12.15 | 0.00 | 0.00 | 183.63 | 0.00 | 479.36 | 675.15 |
| | 4 | 24.31 | 0.00 | 0.00 | 367.35 | 0.00 | 959.08 | 1350.73 |

Table M.14: liblinear-enron-GM5-ALL-ALL-10

| GM5 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 25 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 111.31 | 156.70 | 268.01 |
| | 1 | 6.07 | 0.00 | 0.00 | 91.79 | 0.00 | 582.90 | 680.76 |
| | 2 | 12.15 | 0.00 | 0.00 | 183.63 | 0.00 | 1173.58 | 1369.37 |

Table M.15: liblinear-enron-GM5-ALL-ALL-25

| GM5 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 50 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 111.31 | 470.59 | 581.90 |
| | 1 | 6.07 | 0.00 | 0.00 | 91.79 | 0.00 | 1166.06 | 1263.92 |

Table M.16: liblinear-enron-GM5-ALL-ALL-50

| GM5 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Size (MB) | | | | | | |
| | | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 75 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 111.31 | 875.63 | 986.94 |

Table M.17: liblinear-enron-GM5-ALL-ALL-75

| GB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Size (MB) | | | | | | |
| | | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 5 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 54.31 | 7.61 | 61.91 |
| | 1 | 3.21 | 0.00 | 0.00 | 48.44 | 0.00 | 72.93 | 124.58 |
| | 2 | 6.41 | 0.00 | 0.00 | 96.88 | 0.00 | 137.26 | 240.56 |
| | 4 | 12.82 | 0.00 | 0.00 | 193.78 | 0.00 | 268.59 | 475.20 |
| | 8 | 25.64 | 0.00 | 0.00 | 387.55 | 0.00 | 533.17 | 946.37 |
| | 16 | 51.31 | 0.00 | 0.00 | 775.39 | 0.00 | 1056.93 | 1883.63 |

Table M.18: liblinear-enron-GB3-ALL-ALL-5

| GB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Size (MB) | | | | | | |
| | | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 10 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 54.31 | 22.61 | 76.92 |
| | 1 | 3.21 | 0.00 | 0.00 | 48.44 | 0.00 | 138.64 | 190.29 |
| | 2 | 6.41 | 0.00 | 0.00 | 96.88 | 0.00 | 267.73 | 371.02 |
| | 4 | 12.82 | 0.00 | 0.00 | 193.78 | 0.00 | 521.84 | 728.44 |
| | 8 | 25.64 | 0.00 | 0.00 | 387.55 | 0.00 | 1018.42 | 1431.62 |

Table M.19: liblinear-enron-GB3-ALL-ALL-10

| GB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 25 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 54.31 | 89.19 | 143.49 |
| | 1 | 3.21 | 0.00 | 0.00 | 48.44 | 0.00 | 363.60 | 415.25 |
| | 2 | 6.41 | 0.00 | 0.00 | 96.88 | 0.00 | 667.07 | 770.36 |

Table M.20: liblinear-enron-GB3-ALL-ALL-25

| GB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 50 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 54.31 | 248.36 | 302.66 |
| | 1 | 3.21 | 0.00 | 0.00 | 48.44 | 0.00 | 748.13 | 799.78 |

Table M.21: liblinear-enron-GB3-ALL-ALL-50

| GB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 75 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 54.31 | 453.56 | 507.86 |
| | 1 | 3.21 | 0.00 | 0.00 | 48.44 | 0.00 | 1135.06 | 1186.71 |

Table M.22: liblinear-enron-GB3-ALL-ALL-75

| OSB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 5 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 118.02 | 15.25 | 133.28 |
| | 1 | 12.41 | 0.00 | 0.00 | 187.54 | 0.00 | 259.32 | 459.27 |
| | 2 | 24.82 | 0.00 | 0.00 | 375.10 | 0.00 | 521.04 | 920.95 |
| | 4 | 49.65 | 0.00 | 0.00 | 750.28 | 0.00 | 1031.48 | 1831.40 |

Table M.23: liblinear-enron-OSB3-ALL-ALL-5

# APPENDIX N:
# Liblinear Storage Requirements for the Twitter Short Message Corpus

| GM1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Size (MB) | | | | | | |
| | | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 5 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.20 | 0.09 | 0.29 |
| | 1 | 0.07 | 0.00 | 0.00 | 1.09 | 0.00 | 1.57 | 2.73 |
| | 2 | 0.14 | 0.00 | 0.00 | 2.17 | 0.00 | 3.06 | 5.38 |
| | 4 | 0.29 | 0.00 | 0.00 | 4.35 | 0.00 | 6.05 | 10.69 |
| | 8 | 0.58 | 0.00 | 0.00 | 8.70 | 0.00 | 12.03 | 21.31 |
| | 16 | 1.15 | 0.00 | 0.00 | 17.40 | 0.00 | 23.95 | 42.50 |

Table N.1: liblinear-twitter-GM1-ALL-ALL-5

| GM1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Size (MB) | | | | | | |
| | | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 10 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.20 | 0.24 | 0.44 |
| | 1 | 0.07 | 0.00 | 0.00 | 1.09 | 0.00 | 3.05 | 4.21 |
| | 2 | 0.14 | 0.00 | 0.00 | 2.17 | 0.00 | 5.90 | 8.22 |
| | 4 | 0.29 | 0.00 | 0.00 | 4.35 | 0.00 | 11.61 | 16.24 |
| | 8 | 0.58 | 0.00 | 0.00 | 8.70 | 0.00 | 23.03 | 32.30 |
| | 16 | 1.15 | 0.00 | 0.00 | 17.40 | 0.00 | 45.80 | 64.35 |

Table N.2: liblinear-twitter-GM1-ALL-ALL-10

| GM1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 25 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.20 | 0.85 | 1.05 |
| | 1 | 0.07 | 0.00 | 0.00 | 1.09 | 0.00 | 7.59 | 8.75 |
| | 2 | 0.14 | 0.00 | 0.00 | 2.17 | 0.00 | 14.52 | 16.84 |
| | 4 | 0.29 | 0.00 | 0.00 | 4.35 | 0.00 | 28.38 | 33.01 |
| | 8 | 0.58 | 0.00 | 0.00 | 8.70 | 0.00 | 56.11 | 65.38 |
| | 16 | 1.15 | 0.00 | 0.00 | 17.40 | 0.00 | 111.51 | 130.05 |

Table N.3: liblinear-twitter-GM1-ALL-ALL-25

| GM1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 50 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.20 | 2.14 | 2.34 |
| | 1 | 0.07 | 0.00 | 0.00 | 1.09 | 0.00 | 15.27 | 16.43 |
| | 2 | 0.14 | 0.00 | 0.00 | 2.17 | 0.00 | 28.99 | 31.31 |
| | 4 | 0.29 | 0.00 | 0.00 | 4.35 | 0.00 | 56.44 | 61.07 |
| | 8 | 0.58 | 0.00 | 0.00 | 8.70 | 0.00 | 111.35 | 120.62 |
| | 16 | 1.15 | 0.00 | 0.00 | 17.40 | 0.00 | 221.15 | 239.70 |

Table N.4: liblinear-twitter-GM1-ALL-ALL-50

| GM1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 75 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.20 | 3.69 | 3.89 |
| | 1 | 0.07 | 0.00 | 0.00 | 1.09 | 0.00 | 23.03 | 24.19 |
| | 2 | 0.14 | 0.00 | 0.00 | 2.17 | 0.00 | 43.56 | 45.88 |
| | 4 | 0.29 | 0.00 | 0.00 | 4.35 | 0.00 | 84.60 | 89.23 |
| | 8 | 0.58 | 0.00 | 0.00 | 8.70 | 0.00 | 166.67 | 175.94 |
| | 16 | 1.15 | 0.00 | 0.00 | 17.40 | 0.00 | 330.81 | 349.35 |

Table N.5: liblinear-twitter-GM1-ALL-ALL-75

| GM1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 150 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.20 | 9.35 | 9.54 |
| | 1 | 0.07 | 0.00 | 0.00 | 1.09 | 0.00 | 46.58 | 47.74 |
| | 2 | 0.14 | 0.00 | 0.00 | 2.17 | 0.00 | 87.49 | 89.81 |
| | 4 | 0.29 | 0.00 | 0.00 | 4.35 | 0.00 | 169.36 | 173.99 |
| | 8 | 0.58 | 0.00 | 0.00 | 8.70 | 0.00 | 332.94 | 342.21 |
| | 16 | 1.15 | 0.00 | 0.00 | 17.40 | 0.00 | 660.12 | 678.66 |

Table N.6: liblinear-twitter-GM1-ALL-ALL-150

| GM2 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 5 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 1.13 | 0.25 | 1.38 |
| | 1 | 1.67 | 0.00 | 0.00 | 25.19 | 0.00 | 34.70 | 61.55 |
| | 2 | 3.28 | 0.00 | 0.00 | 49.56 | 0.00 | 68.16 | 121.01 |
| | 4 | 6.56 | 0.00 | 0.00 | 99.13 | 0.00 | 136.24 | 241.93 |
| | 8 | 13.12 | 0.00 | 0.00 | 198.25 | 0.00 | 272.26 | 483.63 |
| | 16 | 26.47 | 0.00 | 0.00 | 400.00 | 0.00 | 549.72 | 976.19 |

Table N.7: liblinear-twitter-GM2-ALL-ALL-5

| GM2 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 10 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 1.13 | 0.81 | 1.94 |
| | 1 | 1.67 | 0.00 | 0.00 | 25.19 | 0.00 | 66.29 | 93.15 |
| | 2 | 3.28 | 0.00 | 0.00 | 49.56 | 0.00 | 130.21 | 183.05 |
| | 4 | 6.56 | 0.00 | 0.00 | 99.13 | 0.00 | 260.24 | 365.93 |
| | 8 | 13.12 | 0.00 | 0.00 | 198.25 | 0.00 | 519.95 | 731.33 |
| | 16 | 26.47 | 0.00 | 0.00 | 400.00 | 0.00 | 1049.50 | 1475.97 |

Table N.8: liblinear-twitter-GM2-ALL-ALL-10

| GM2 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 25 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 1.13 | 3.27 | 4.40 |
| | 1 | 1.67 | 0.00 | 0.00 | 25.19 | 0.00 | 161.17 | 188.03 |
| | 2 | 3.28 | 0.00 | 0.00 | 49.56 | 0.00 | 316.40 | 369.24 |
| | 4 | 6.56 | 0.00 | 0.00 | 99.13 | 0.00 | 632.32 | 738.01 |

Table N.9: liblinear-twitter-GM2-ALL-ALL-25

| GM2 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 50 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 1.13 | 8.87 | 10.00 |
| | 1 | 1.67 | 0.00 | 0.00 | 25.19 | 0.00 | 319.44 | 346.29 |
| | 2 | 3.28 | 0.00 | 0.00 | 49.56 | 0.00 | 626.91 | 679.75 |

Table N.10: liblinear-twitter-GM2-ALL-ALL-50

| GM2 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 75 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 1.13 | 15.62 | 16.75 |
| | 1 | 1.67 | 0.00 | 0.00 | 25.19 | 0.00 | 477.78 | 504.64 |
| | 2 | 3.28 | 0.00 | 0.00 | 49.56 | 0.00 | 937.67 | 990.51 |

Table N.11: liblinear-twitter-GM2-ALL-ALL-75

| GM2 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 150 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 1.13 | 41.99 | 43.12 |
| | 1 | 1.67 | 0.00 | 0.00 | 25.19 | 0.00 | 953.13 | 979.99 |

Table N.12: liblinear-twitter-GM2-ALL-ALL-150

| GM5 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 5 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 3.08 | 0.31 | 3.40 |
| | 1 | 6.07 | 0.00 | 0.00 | 91.79 | 0.00 | 125.71 | 223.57 |
| | 2 | 12.15 | 0.00 | 0.00 | 183.63 | 0.00 | 251.37 | 447.15 |

Table N.13: liblinear-twitter-GM5-ALL-ALL-5

| GM5 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 10 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 3.08 | 1.24 | 4.32 |
| | 1 | 6.07 | 0.00 | 0.00 | 91.79 | 0.00 | 240.86 | 338.71 |
| | 2 | 12.15 | 0.00 | 0.00 | 183.63 | 0.00 | 482.06 | 677.85 |

Table N.14: liblinear-twitter-GM5-ALL-ALL-10

| GM5 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 25 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 3.08 | 6.69 | 9.77 |
| | 1 | 6.07 | 0.00 | 0.00 | 91.79 | 0.00 | 585.06 | 682.92 |

Table N.15: liblinear-twitter-GM5-ALL-ALL-25

| GM5 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 50 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 3.08 | 21.34 | 24.42 |

Table N.16: liblinear-twitter-GM5-ALL-ALL-50

| GM5 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 75 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 3.08 | 39.19 | 42.27 |

Table N.17: liblinear-twitter-GM5-ALL-ALL-75

| GM5 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 150 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 3.08 | 111.25 | 114.33 |

Table N.18: liblinear-twitter-GM5-ALL-ALL-150

| GB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Size (MB) | | | | | | |
| | | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 5 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 3.18 | 0.78 | 3.96 |
| | 1 | 3.21 | 0.00 | 0.00 | 48.44 | 0.00 | 67.05 | 118.69 |
| | 2 | 6.41 | 0.00 | 0.00 | 96.88 | 0.00 | 133.78 | 237.07 |
| | 4 | 12.82 | 0.00 | 0.00 | 193.78 | 0.00 | 266.81 | 473.41 |
| | 8 | 25.64 | 0.00 | 0.00 | 387.55 | 0.00 | 533.35 | 946.54 |
| | 16 | 51.31 | 0.00 | 0.00 | 775.39 | 0.00 | 1066.46 | 1893.16 |

Table N.19: liblinear-twitter-GB3-ALL-ALL-5

| GB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Size (MB) | | | | | | |
| | | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 10 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 3.18 | 2.56 | 5.74 |
| | 1 | 3.21 | 0.00 | 0.00 | 48.44 | 0.00 | 128.85 | 180.50 |
| | 2 | 6.41 | 0.00 | 0.00 | 96.88 | 0.00 | 256.35 | 359.64 |
| | 4 | 12.82 | 0.00 | 0.00 | 193.78 | 0.00 | 510.53 | 717.13 |
| | 8 | 25.64 | 0.00 | 0.00 | 387.55 | 0.00 | 1018.56 | 1431.76 |

Table N.20: liblinear-twitter-GB3-ALL-ALL-10

| GB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Size (MB) | | | | | | |
| | | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 25 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 3.18 | 10.52 | 13.70 |
| | 1 | 3.21 | 0.00 | 0.00 | 48.44 | 0.00 | 314.97 | 366.62 |
| | 2 | 6.41 | 0.00 | 0.00 | 96.88 | 0.00 | 625.60 | 728.89 |

Table N.21: liblinear-twitter-GB3-ALL-ALL-25

| GB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 50 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 3.18 | 28.75 | 31.93 |
| | 1 | 3.21 | 0.00 | 0.00 | 48.44 | 0.00 | 629.59 | 681.23 |

Table N.22: liblinear-twitter-GB3-ALL-ALL-50

| GB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 75 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 3.18 | 51.29 | 54.47 |
| | 1 | 3.21 | 0.00 | 0.00 | 48.44 | 0.00 | 947.75 | 999.39 |

Table N.23: liblinear-twitter-GB3-ALL-ALL-75

| GB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 150 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 3.18 | 132.47 | 135.65 |

Table N.24: liblinear-twitter-GB3-ALL-ALL-150

| OSB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 5 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 7.30 | 1.62 | 8.92 |
| | 1 | 12.41 | 0.00 | 0.00 | 187.54 | 0.00 | 258.70 | 458.66 |
| | 2 | 24.82 | 0.00 | 0.00 | 375.10 | 0.00 | 516.75 | 916.66 |
| | 4 | 49.65 | 0.00 | 0.00 | 750.28 | 0.00 | 1032.46 | 1832.38 |

Table N.25: liblinear-twitter-OSB3-ALL-ALL-5

| OSB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 10 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 7.30 | 5.33 | 12.63 |
| | 1 | 12.41 | 0.00 | 0.00 | 187.54 | 0.00 | 496.13 | 696.09 |
| | 2 | 24.82 | 0.00 | 0.00 | 375.10 | 0.00 | 987.43 | 1387.34 |

Table N.26: liblinear-twitter-OSB3-ALL-ALL-10

| OSB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 25 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 7.30 | 22.51 | 29.81 |

Table N.27: liblinear-twitter-OSB3-ALL-ALL-25

| OSB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 50 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 7.30 | 62.19 | 69.49 |

Table N.28: liblinear-twitter-OSB3-ALL-ALL-50

| OSB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 75 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 7.30 | 109.61 | 116.92 |

Table N.29: liblinear-twitter-OSB3-ALL-ALL-75

| OSB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 150 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 7.30 | 288.83 | 296.14 |

Table N.30: liblinear-twitter-OSB3-ALL-ALL-150

# APPENDIX O:
# Naive Bayes Storage Requirements for the ENRON Email Corpus

| GM1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 5 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 1.40 | 0.21 | 1.60 |
| | 1 | 0.07 | 0.27 | 1.09 | 1.09 | 0.00 | 0.33 | 2.85 |
| | 2 | 0.14 | 0.54 | 2.17 | 2.17 | 0.00 | 0.33 | 5.37 |
| | 4 | 0.29 | 1.09 | 4.35 | 4.35 | 0.00 | 0.33 | 10.40 |
| | 8 | 0.58 | 2.17 | 8.70 | 8.70 | 0.00 | 0.33 | 20.48 |
| | 16 | 1.15 | 4.35 | 17.40 | 17.40 | 0.00 | 0.33 | 40.62 |

Table O.1: nb-enron-GM1-ALL-ALL-5

| GM1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 10 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 1.40 | 0.41 | 1.81 |
| | 1 | 0.07 | 0.27 | 1.09 | 1.09 | 0.00 | 0.66 | 3.18 |
| | 2 | 0.14 | 0.54 | 2.17 | 2.17 | 0.00 | 0.66 | 5.70 |
| | 4 | 0.29 | 1.09 | 4.35 | 4.35 | 0.00 | 0.66 | 10.73 |
| | 8 | 0.58 | 2.17 | 8.70 | 8.70 | 0.00 | 0.66 | 20.81 |
| | 16 | 1.15 | 4.35 | 17.40 | 17.40 | 0.00 | 0.66 | 40.95 |

Table O.2: nb-enron-GM1-ALL-ALL-10

| GM1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 25 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 1.40 | 1.03 | 2.43 |
| | 1 | 0.07 | 0.27 | 1.09 | 1.09 | 0.00 | 1.64 | 4.16 |
| | 2 | 0.14 | 0.54 | 2.17 | 2.17 | 0.00 | 1.65 | 6.69 |
| | 4 | 0.29 | 1.09 | 4.35 | 4.35 | 0.00 | 1.65 | 11.72 |
| | 8 | 0.58 | 2.17 | 8.70 | 8.70 | 0.00 | 1.65 | 21.80 |
| | 16 | 1.15 | 4.35 | 17.40 | 17.40 | 0.00 | 1.65 | 41.94 |

Table O.3: nb-enron-GM1-ALL-ALL-25

| GM1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 50 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 1.40 | 2.07 | 3.46 |
| | 1 | 0.07 | 0.27 | 1.09 | 1.09 | 0.00 | 3.29 | 5.81 |
| | 2 | 0.14 | 0.54 | 2.17 | 2.17 | 0.00 | 3.30 | 8.34 |
| | 4 | 0.29 | 1.09 | 4.35 | 4.35 | 0.00 | 3.31 | 13.38 |
| | 8 | 0.58 | 2.17 | 8.70 | 8.70 | 0.00 | 3.31 | 23.45 |
| | 16 | 1.15 | 4.35 | 17.40 | 17.40 | 0.00 | 3.31 | 43.60 |

Table O.4: nb-enron-GM1-ALL-ALL-50

| GM1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Size (MB) | | | | | | |
| | | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 75 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 1.40 | 3.10 | 4.50 |
| | 1 | 0.07 | 0.27 | 1.09 | 1.09 | 0.00 | 4.93 | 7.45 |
| | 2 | 0.14 | 0.54 | 2.17 | 2.17 | 0.00 | 4.95 | 9.99 |
| | 4 | 0.29 | 1.09 | 4.35 | 4.35 | 0.00 | 4.96 | 15.03 |
| | 8 | 0.58 | 2.17 | 8.70 | 8.70 | 0.00 | 4.96 | 25.11 |
| | 16 | 1.15 | 4.35 | 17.40 | 17.40 | 0.00 | 4.96 | 45.25 |

Table O.5: nb-enron-GM1-ALL-ALL-75

| GM1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Size (MB) | | | | | | |
| | | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 150 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 1.40 | 6.18 | 7.58 |
| | 1 | 0.07 | 0.27 | 1.09 | 1.09 | 0.00 | 9.87 | 12.39 |
| | 2 | 0.14 | 0.54 | 2.17 | 2.17 | 0.00 | 9.90 | 14.94 |
| | 4 | 0.29 | 1.09 | 4.35 | 4.35 | 0.00 | 9.92 | 19.99 |
| | 8 | 0.58 | 2.17 | 8.70 | 8.70 | 0.00 | 9.93 | 30.07 |
| | 16 | 1.15 | 4.35 | 17.40 | 17.40 | 0.00 | 9.93 | 50.22 |

Table O.6: nb-enron-GM1-ALL-ALL-150

| GM2 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 5 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 16.61 | 0.87 | 17.48 |
| | 1 | 1.67 | 6.30 | 25.19 | 25.19 | 0.00 | 0.33 | 58.67 |
| | 2 | 3.28 | 12.39 | 49.56 | 49.56 | 0.00 | 0.33 | 115.12 |
| | 4 | 6.56 | 24.78 | 99.13 | 99.13 | 0.00 | 0.33 | 229.93 |
| | 8 | 13.12 | 49.56 | 198.25 | 198.25 | 0.00 | 0.33 | 459.52 |
| | 16 | 26.47 | 100.00 | 400.00 | 400.00 | 0.00 | 0.33 | 926.79 |

Table O.7: nb-enron-GM2-ALL-ALL-5

| GM2 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 10 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 16.61 | 1.73 | 18.34 |
| | 1 | 1.67 | 6.30 | 25.19 | 25.19 | 0.00 | 0.65 | 58.99 |
| | 2 | 3.28 | 12.39 | 49.56 | 49.56 | 0.00 | 0.65 | 115.45 |
| | 4 | 6.56 | 24.78 | 99.13 | 99.13 | 0.00 | 0.65 | 230.26 |
| | 8 | 13.12 | 49.56 | 198.25 | 198.25 | 0.00 | 0.65 | 459.84 |
| | 16 | 26.47 | 100.00 | 400.00 | 400.00 | 0.00 | 0.65 | 927.12 |

Table O.8: nb-enron-GM2-ALL-ALL-10

| GM2 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Size (MB) | | | | | | |
| | | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 25 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 16.61 | 4.38 | 20.99 |
| | 1 | 1.67 | 6.30 | 25.19 | 25.19 | 0.00 | 1.63 | 59.97 |
| | 2 | 3.28 | 12.39 | 49.56 | 49.56 | 0.00 | 1.63 | 116.42 |
| | 4 | 6.56 | 24.78 | 99.13 | 99.13 | 0.00 | 1.63 | 231.24 |
| | 8 | 13.12 | 49.56 | 198.25 | 198.25 | 0.00 | 1.63 | 460.82 |
| | 16 | 26.47 | 100.00 | 400.00 | 400.00 | 0.00 | 1.63 | 928.09 |

Table O.9: nb-enron-GM2-ALL-ALL-25

| GM2 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Size (MB) | | | | | | |
| | | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 50 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 16.61 | 8.69 | 25.30 |
| | 1 | 1.67 | 6.30 | 25.19 | 25.19 | 0.00 | 3.25 | 61.59 |
| | 2 | 3.28 | 12.39 | 49.56 | 49.56 | 0.00 | 3.25 | 118.05 |
| | 4 | 6.56 | 24.78 | 99.13 | 99.13 | 0.00 | 3.26 | 232.86 |
| | 8 | 13.12 | 49.56 | 198.25 | 198.25 | 0.00 | 3.26 | 462.45 |
| | 16 | 26.47 | 100.00 | 400.00 | 400.00 | 0.00 | 3.26 | 929.72 |

Table O.10: nb-enron-GM2-ALL-ALL-50

| GM2 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 75 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 16.61 | 13.21 | 29.82 |
| | 1 | 1.67 | 6.30 | 25.19 | 25.19 | 0.00 | 4.88 | 63.22 |
| | 2 | 3.28 | 12.39 | 49.56 | 49.56 | 0.00 | 4.88 | 119.68 |
| | 4 | 6.56 | 24.78 | 99.13 | 99.13 | 0.00 | 4.88 | 234.49 |
| | 8 | 13.12 | 49.56 | 198.25 | 198.25 | 0.00 | 4.88 | 464.07 |
| | 16 | 26.47 | 100.00 | 400.00 | 400.00 | 0.00 | 4.88 | 931.35 |

Table O.11: nb-enron-GM2-ALL-ALL-75

| GM2 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 150 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 16.61 | 25.99 | 42.60 |
| | 1 | 1.67 | 6.30 | 25.19 | 25.19 | 0.00 | 9.75 | 68.10 |
| | 2 | 3.28 | 12.39 | 49.56 | 49.56 | 0.00 | 9.76 | 124.56 |
| | 4 | 6.56 | 24.78 | 99.13 | 99.13 | 0.00 | 9.77 | 239.37 |
| | 8 | 13.12 | 49.56 | 198.25 | 198.25 | 0.00 | 9.77 | 468.96 |
| | 16 | 26.47 | 100.00 | 400.00 | 400.00 | 0.00 | 9.77 | 936.23 |

Table O.12: nb-enron-GM2-ALL-ALL-150

| GM5 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 5 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 111.31 | 1.63 | 112.94 |
| | 1 | 6.07 | 22.95 | 91.79 | 91.79 | 0.00 | 0.32 | 212.91 |
| | 2 | 12.15 | 45.91 | 183.63 | 183.63 | 0.00 | 0.32 | 425.65 |
| | 4 | 24.31 | 91.84 | 367.35 | 367.35 | 0.00 | 0.32 | 851.16 |
| | 8 | 48.70 | 184.00 | 735.99 | 735.99 | 0.00 | 0.32 | 1704.99 |
| | 16 | 97.43 | 368.11 | 1472.43 | 1472.43 | 0.00 | 0.32 | 3410.72 |

Table O.13: nb-enron-GM5-ALL-ALL-5

| GM5 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 10 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 111.31 | 3.27 | 114.58 |
| | 1 | 6.07 | 22.95 | 91.79 | 91.79 | 0.00 | 0.63 | 213.23 |
| | 2 | 12.15 | 45.91 | 183.63 | 183.63 | 0.00 | 0.63 | 425.96 |
| | 4 | 24.31 | 91.84 | 367.35 | 367.35 | 0.00 | 0.63 | 851.48 |
| | 8 | 48.70 | 184.00 | 735.99 | 735.99 | 0.00 | 0.63 | 1705.31 |
| | 16 | 97.43 | 368.11 | 1472.43 | 1472.43 | 0.00 | 0.63 | 3411.04 |

Table O.14: nb-enron-GM5-ALL-ALL-10

| GM5 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 25 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 111.31 | 7.99 | 119.30 |
| | 1 | 6.07 | 22.95 | 91.79 | 91.79 | 0.00 | 1.58 | 214.18 |
| | 2 | 12.15 | 45.91 | 183.63 | 183.63 | 0.00 | 1.58 | 426.91 |
| | 4 | 24.31 | 91.84 | 367.35 | 367.35 | 0.00 | 1.58 | 852.43 |
| | 8 | 48.70 | 184.00 | 735.99 | 735.99 | 0.00 | 1.58 | 1706.26 |
| | 16 | 97.43 | 368.11 | 1472.43 | 1472.43 | 0.00 | 1.59 | 3411.99 |

Table O.15: nb-enron-GM5-ALL-ALL-25

| GM5 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 50 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 111.31 | 16.13 | 127.44 |
| | 1 | 6.07 | 22.95 | 91.79 | 91.79 | 0.00 | 3.17 | 215.76 |
| | 2 | 12.15 | 45.91 | 183.63 | 183.63 | 0.00 | 3.17 | 428.50 |
| | 4 | 24.31 | 91.84 | 367.35 | 367.35 | 0.00 | 3.17 | 854.01 |
| | 8 | 48.70 | 184.00 | 735.99 | 735.99 | 0.00 | 3.17 | 1707.85 |
| | 16 | 97.43 | 368.11 | 1472.43 | 1472.43 | 0.00 | 3.17 | 3413.58 |

Table O.16: nb-enron-GM5-ALL-ALL-50

| GM5 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 75 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 111.31 | 24.39 | 135.70 |
| | 1 | 6.07 | 22.95 | 91.79 | 91.79 | 0.00 | 4.75 | 217.34 |
| | 2 | 12.15 | 45.91 | 183.63 | 183.63 | 0.00 | 4.75 | 430.08 |
| | 4 | 24.31 | 91.84 | 367.35 | 367.35 | 0.00 | 4.75 | 855.60 |
| | 8 | 48.70 | 184.00 | 735.99 | 735.99 | 0.00 | 4.75 | 1709.43 |
| | 16 | 97.43 | 368.11 | 1472.43 | 1472.43 | 0.00 | 4.75 | 3415.16 |

Table O.17: nb-enron-GM5-ALL-ALL-75

| GM5 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 150 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 111.31 | 48.92 | 160.23 |
| | 1 | 6.07 | 22.95 | 91.79 | 91.79 | 0.00 | 9.51 | 222.10 |
| | 2 | 12.15 | 45.91 | 183.63 | 183.63 | 0.00 | 9.51 | 434.84 |
| | 4 | 24.31 | 91.84 | 367.35 | 367.35 | 0.00 | 9.51 | 860.35 |
| | 8 | 48.70 | 184.00 | 735.99 | 735.99 | 0.00 | 9.51 | 1714.19 |
| | 16 | 97.43 | 368.11 | 1472.43 | 1472.43 | 0.00 | 9.51 | 3419.91 |

Table O.18: nb-enron-GM5-ALL-ALL-150

| GB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 5 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 54.31 | 2.67 | 56.98 |
| | 1 | 3.21 | 12.11 | 48.44 | 48.44 | 0.00 | 4.22 | 116.41 |
| | 2 | 6.41 | 24.22 | 96.88 | 96.88 | 0.00 | 4.24 | 228.64 |
| | 4 | 12.82 | 48.44 | 193.78 | 193.78 | 0.00 | 4.26 | 453.08 |
| | 8 | 25.64 | 96.89 | 387.55 | 387.55 | 0.00 | 4.27 | 901.90 |
| | 16 | 51.31 | 193.85 | 775.39 | 775.39 | 0.00 | 4.27 | 1800.20 |

Table O.19: nb-enron-GB3-ALL-ALL-5

| GB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 10 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 54.31 | 5.31 | 59.61 |
| | 1 | 3.21 | 12.11 | 48.44 | 48.44 | 0.00 | 8.43 | 120.63 |
| | 2 | 6.41 | 24.22 | 96.88 | 96.88 | 0.00 | 8.49 | 232.88 |
| | 4 | 12.82 | 48.44 | 193.78 | 193.78 | 0.00 | 8.52 | 457.34 |
| | 8 | 25.64 | 96.89 | 387.55 | 387.55 | 0.00 | 8.53 | 906.17 |
| | 16 | 51.31 | 193.85 | 775.39 | 775.39 | 0.00 | 8.54 | 1804.47 |

Table O.20: nb-enron-GB3-ALL-ALL-10

| GB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 25 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 54.31 | 13.42 | 67.73 |
| | 1 | 3.21 | 12.11 | 48.44 | 48.44 | 0.00 | 21.08 | 133.28 |
| | 2 | 6.41 | 24.22 | 96.88 | 96.88 | 0.00 | 21.22 | 245.61 |
| | 4 | 12.82 | 48.44 | 193.78 | 193.78 | 0.00 | 21.29 | 470.12 |
| | 8 | 25.64 | 96.89 | 387.55 | 387.55 | 0.00 | 21.33 | 918.97 |
| | 16 | 51.31 | 193.85 | 775.39 | 775.39 | 0.00 | 21.34 | 1817.28 |

Table O.21: nb-enron-GB3-ALL-ALL-25

| GB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 50 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 54.31 | 26.67 | 80.97 |
| | 1 | 3.21 | 12.11 | 48.44 | 48.44 | 0.00 | 42.16 | 154.36 |
| | 2 | 6.41 | 24.22 | 96.88 | 96.88 | 0.00 | 42.44 | 266.83 |
| | 4 | 12.82 | 48.44 | 193.78 | 193.78 | 0.00 | 42.58 | 491.40 |
| | 8 | 25.64 | 96.89 | 387.55 | 387.55 | 0.00 | 42.65 | 940.29 |
| | 16 | 51.31 | 193.85 | 775.39 | 775.39 | 0.00 | 42.68 | 1838.62 |

Table O.22: nb-enron-GB3-ALL-ALL-50

| GB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 75 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 54.31 | 40.22 | 94.53 |
| | 1 | 3.21 | 12.11 | 48.44 | 48.44 | 0.00 | 63.24 | 175.43 |
| | 2 | 6.41 | 24.22 | 96.88 | 96.88 | 0.00 | 63.66 | 288.05 |
| | 4 | 12.82 | 48.44 | 193.78 | 193.78 | 0.00 | 63.87 | 512.69 |
| | 8 | 25.64 | 96.89 | 387.55 | 387.55 | 0.00 | 63.98 | 961.62 |
| | 16 | 51.31 | 193.85 | 775.39 | 775.39 | 0.00 | 64.03 | 1859.96 |

Table O.23: nb-enron-GB3-ALL-ALL-75

| GB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 150 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 54.31 | 79.47 | 133.78 |
| | 1 | 3.21 | 12.11 | 48.44 | 48.44 | 0.00 | 126.47 | 238.67 |
| | 2 | 6.41 | 24.22 | 96.88 | 96.88 | 0.00 | 127.32 | 351.71 |
| | 4 | 12.82 | 48.44 | 193.78 | 193.78 | 0.00 | 127.73 | 576.56 |
| | 8 | 25.64 | 96.89 | 387.55 | 387.55 | 0.00 | 127.96 | 1025.59 |
| | 16 | 51.31 | 193.85 | 775.39 | 775.39 | 0.00 | 128.05 | 1923.99 |

Table O.24: nb-enron-GB3-ALL-ALL-150

| OSB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Size (MB) | | | | | | |
| | | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 5 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 118.02 | 5.33 | 123.35 |
| | 1 | 12.41 | 46.89 | 187.54 | 187.54 | 0.00 | 8.53 | 442.91 |
| | 2 | 24.82 | 93.77 | 375.10 | 375.10 | 0.00 | 8.54 | 877.33 |
| | 4 | 49.65 | 187.57 | 750.28 | 750.28 | 0.00 | 8.55 | 1746.33 |
| | 8 | 99.29 | 375.14 | 1500.55 | 1500.55 | 0.00 | 8.55 | 3484.08 |
| | 16 | 198.81 | 751.03 | 3004.14 | 3004.14 | 0.00 | 8.55 | 6966.67 |

Table O.25: nb-enron-OSB3-ALL-ALL-5

| OSB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Size (MB) | | | | | | |
| | | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 10 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 118.02 | 10.62 | 128.64 |
| | 1 | 12.41 | 46.89 | 187.54 | 187.54 | 0.00 | 17.06 | 451.44 |
| | 2 | 24.82 | 93.77 | 375.10 | 375.10 | 0.00 | 17.09 | 885.87 |
| | 4 | 49.65 | 187.57 | 750.28 | 750.28 | 0.00 | 17.10 | 1754.88 |
| | 8 | 99.29 | 375.14 | 1500.55 | 1500.55 | 0.00 | 17.11 | 3492.64 |
| | 16 | 198.81 | 751.03 | 3004.14 | 3004.14 | 0.00 | 11.94 | 6970.05 |

Table O.26: nb-enron-OSB3-ALL-ALL-10

| OSB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Size (MB) | | | | | | |
| | | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 25 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 118.02 | 26.84 | 144.86 |
| | 1 | 12.41 | 46.89 | 187.54 | 187.54 | 0.00 | 42.64 | 477.03 |
| | 2 | 24.82 | 93.77 | 375.10 | 375.10 | 0.00 | 42.72 | 911.50 |
| | 4 | 49.65 | 187.57 | 750.28 | 750.28 | 0.00 | 42.75 | 1780.53 |
| | 8 | 99.29 | 375.14 | 1500.55 | 1500.55 | 0.00 | 42.77 | 3518.29 |

Table O.27: nb-enron-OSB3-ALL-ALL-25

| OSB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Size (MB) | | | | | | |
| | | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 50 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 118.02 | 53.35 | 171.37 |
| | 1 | 12.41 | 46.89 | 187.54 | 187.54 | 0.00 | 85.28 | 519.67 |
| | 2 | 24.82 | 93.77 | 375.10 | 375.10 | 0.00 | 85.43 | 954.22 |
| | 4 | 49.65 | 187.57 | 750.28 | 750.28 | 0.00 | 85.50 | 1823.28 |

Table O.28: nb-enron-OSB3-ALL-ALL-50

| OSB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Size (MB) | | | | | | |
| | | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 75 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 118.02 | 80.47 | 198.49 |
| | 1 | 12.41 | 46.89 | 187.54 | 187.54 | 0.00 | 127.92 | 562.31 |
| | 2 | 24.82 | 93.77 | 375.10 | 375.10 | 0.00 | 128.15 | 996.94 |
| | 4 | 49.65 | 187.57 | 750.28 | 750.28 | 0.00 | 128.25 | 1866.03 |
| | 8 | 99.29 | 375.14 | 1500.55 | 1500.55 | 0.00 | 128.30 | 3603.82 |

Table O.29: nb-enron-OSB3-ALL-ALL-75

| OSB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Size (MB) | | | | | | |
| | | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 150 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 118.02 | 158.91 | 276.94 |
| | 1 | 12.41 | 46.89 | 187.54 | 187.54 | 0.00 | 255.84 | 690.22 |
| | 2 | 24.82 | 93.77 | 375.10 | 375.10 | 0.00 | 256.30 | 1125.09 |
| | 4 | 49.65 | 187.57 | 750.28 | 750.28 | 0.00 | 256.51 | 1994.29 |
| | 8 | 99.29 | 375.14 | 1500.55 | 1500.55 | 0.00 | 256.59 | 3732.12 |

Table O.30: nb-enron-OSB3-ALL-ALL-150

THIS PAGE INTENTIONALLY LEFT BLANK

# APPENDIX P:
# Naive Bayes Storage Requirements for the Twitter Short Message Corpus

| GM1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 5 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.20 | 0.02 | 0.22 |
| | 1 | 0.07 | 0.27 | 1.09 | 1.09 | 0.00 | 0.04 | 2.56 |
| | 2 | 0.14 | 0.54 | 2.17 | 2.17 | 0.00 | 0.04 | 5.08 |
| | 4 | 0.29 | 1.09 | 4.35 | 4.35 | 0.00 | 0.04 | 10.11 |
| | 8 | 0.58 | 2.17 | 8.70 | 8.70 | 0.00 | 0.04 | 20.18 |
| | 16 | 1.15 | 4.35 | 17.40 | 17.40 | 0.00 | 0.04 | 40.33 |

Table P.1: nb-twitter-GM1-ALL-ALL-5

| GM1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 10 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.20 | 0.05 | 0.25 |
| | 1 | 0.07 | 0.27 | 1.09 | 1.09 | 0.00 | 0.08 | 2.60 |
| | 2 | 0.14 | 0.54 | 2.17 | 2.17 | 0.00 | 0.08 | 5.12 |
| | 4 | 0.29 | 1.09 | 4.35 | 4.35 | 0.00 | 0.08 | 10.15 |
| | 8 | 0.58 | 2.17 | 8.70 | 8.70 | 0.00 | 0.08 | 20.22 |
| | 16 | 1.15 | 4.35 | 17.40 | 17.40 | 0.00 | 0.08 | 40.37 |

Table P.2: nb-twitter-GM1-ALL-ALL-10

| GM1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 25 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.20 | 0.12 | 0.32 |
| | 1 | 0.07 | 0.27 | 1.09 | 1.09 | 0.00 | 0.20 | 2.72 |
| | 2 | 0.14 | 0.54 | 2.17 | 2.17 | 0.00 | 0.20 | 5.23 |
| | 4 | 0.29 | 1.09 | 4.35 | 4.35 | 0.00 | 0.20 | 10.27 |
| | 8 | 0.58 | 2.17 | 8.70 | 8.70 | 0.00 | 0.20 | 20.34 |
| | 16 | 1.15 | 4.35 | 17.40 | 17.40 | 0.00 | 0.20 | 40.49 |

Table P.3: nb-twitter-GM1-ALL-ALL-25

| GM1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 50 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.20 | 0.24 | 0.43 |
| | 1 | 0.07 | 0.27 | 1.09 | 1.09 | 0.00 | 0.40 | 2.91 |
| | 2 | 0.14 | 0.54 | 2.17 | 2.17 | 0.00 | 0.40 | 5.43 |
| | 4 | 0.29 | 1.09 | 4.35 | 4.35 | 0.00 | 0.39 | 10.47 |
| | 8 | 0.58 | 2.17 | 8.70 | 8.70 | 0.00 | 0.40 | 20.54 |
| | 16 | 1.15 | 4.35 | 17.40 | 17.40 | 0.00 | 0.40 | 40.69 |

Table P.4: nb-twitter-GM1-ALL-ALL-50

| GM1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 75 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.20 | 0.36 | 0.55 |
| | 1 | 0.07 | 0.27 | 1.09 | 1.09 | 0.00 | 0.60 | 3.11 |
| | 2 | 0.14 | 0.54 | 2.17 | 2.17 | 0.00 | 0.60 | 5.63 |
| | 4 | 0.29 | 1.09 | 4.35 | 4.35 | 0.00 | 0.60 | 10.67 |
| | 8 | 0.58 | 2.17 | 8.70 | 8.70 | 0.00 | 0.59 | 20.74 |
| | 16 | 1.15 | 4.35 | 17.40 | 17.40 | 0.00 | 0.59 | 40.88 |

Table P.5: nb-twitter-GM1-ALL-ALL-75

| GM1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 150 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.20 | 0.71 | 0.91 |
| | 1 | 0.07 | 0.27 | 1.09 | 1.09 | 0.00 | 1.18 | 3.70 |
| | 2 | 0.14 | 0.54 | 2.17 | 2.17 | 0.00 | 1.19 | 6.23 |
| | 4 | 0.29 | 1.09 | 4.35 | 4.35 | 0.00 | 1.19 | 11.26 |
| | 8 | 0.58 | 2.17 | 8.70 | 8.70 | 0.00 | 1.19 | 21.33 |
| | 16 | 1.15 | 4.35 | 17.40 | 17.40 | 0.00 | 1.18 | 41.47 |

Table P.6: nb-twitter-GM1-ALL-ALL-150

| GM2 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 5 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 1.13 | 0.04 | 1.17 |
| | 1 | 1.67 | 6.30 | 25.19 | 25.19 | 0.00 | 0.04 | 58.38 |
| | 2 | 3.28 | 12.39 | 49.56 | 49.56 | 0.00 | 0.04 | 114.84 |
| | 4 | 6.56 | 24.78 | 99.13 | 99.13 | 0.00 | 0.04 | 229.65 |
| | 8 | 13.12 | 49.56 | 198.25 | 198.25 | 0.00 | 0.04 | 459.23 |
| | 16 | 26.47 | 100.00 | 400.00 | 400.00 | 0.00 | 0.04 | 926.51 |

Table P.7: nb-twitter-GM2-ALL-ALL-5

| GM2 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 10 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 1.13 | 0.09 | 1.22 |
| | 1 | 1.67 | 6.30 | 25.19 | 25.19 | 0.00 | 0.08 | 58.42 |
| | 2 | 3.28 | 12.39 | 49.56 | 49.56 | 0.00 | 0.08 | 114.87 |
| | 4 | 6.56 | 24.78 | 99.13 | 99.13 | 0.00 | 0.08 | 229.68 |
| | 8 | 13.12 | 49.56 | 198.25 | 198.25 | 0.00 | 0.08 | 459.27 |
| | 16 | 26.47 | 100.00 | 400.00 | 400.00 | 0.00 | 0.08 | 926.54 |

Table P.8: nb-twitter-GM2-ALL-ALL-10

| GM2 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 25 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 1.13 | 0.22 | 1.35 |
| | 1 | 1.67 | 6.30 | 25.19 | 25.19 | 0.00 | 0.19 | 58.53 |
| | 2 | 3.28 | 12.39 | 49.56 | 49.56 | 0.00 | 0.19 | 114.99 |
| | 4 | 6.56 | 24.78 | 99.13 | 99.13 | 0.00 | 0.19 | 229.80 |
| | 8 | 13.12 | 49.56 | 198.25 | 198.25 | 0.00 | 0.19 | 459.38 |
| | 16 | 26.47 | 100.00 | 400.00 | 400.00 | 0.00 | 0.19 | 926.66 |

Table P.9: nb-twitter-GM2-ALL-ALL-25

| GM2 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 50 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 1.13 | 0.44 | 1.57 |
| | 1 | 1.67 | 6.30 | 25.19 | 25.19 | 0.00 | 0.38 | 58.72 |
| | 2 | 3.28 | 12.39 | 49.56 | 49.56 | 0.00 | 0.38 | 115.18 |
| | 4 | 6.56 | 24.78 | 99.13 | 99.13 | 0.00 | 0.38 | 229.99 |
| | 8 | 13.12 | 49.56 | 198.25 | 198.25 | 0.00 | 0.38 | 459.57 |
| | 16 | 26.47 | 100.00 | 400.00 | 400.00 | 0.00 | 0.38 | 926.85 |

Table P.10: nb-twitter-GM2-ALL-ALL-50

| GM2 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Size (MB) | | | | | | |
| | | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 75 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 1.13 | 0.66 | 1.79 |
| | 1 | 1.67 | 6.30 | 25.19 | 25.19 | 0.00 | 0.57 | 58.91 |
| | 2 | 3.28 | 12.39 | 49.56 | 49.56 | 0.00 | 0.58 | 115.37 |
| | 4 | 6.56 | 24.78 | 99.13 | 99.13 | 0.00 | 0.58 | 230.19 |
| | 8 | 13.12 | 49.56 | 198.25 | 198.25 | 0.00 | 0.58 | 459.77 |
| | 16 | 26.47 | 100.00 | 400.00 | 400.00 | 0.00 | 0.58 | 927.05 |

Table P.11: nb-twitter-GM2-ALL-ALL-75

| GM2 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Size (MB) | | | | | | |
| | | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 150 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 1.13 | 1.33 | 2.46 |
| | 1 | 1.67 | 6.30 | 25.19 | 25.19 | 0.00 | 1.15 | 59.49 |
| | 2 | 3.28 | 12.39 | 49.56 | 49.56 | 0.00 | 1.15 | 115.95 |
| | 4 | 6.56 | 24.78 | 99.13 | 99.13 | 0.00 | 1.15 | 230.76 |
| | 8 | 13.12 | 49.56 | 198.25 | 198.25 | 0.00 | 1.15 | 460.34 |
| | 16 | 26.47 | 100.00 | 400.00 | 400.00 | 0.00 | 1.15 | 927.61 |

Table P.12: nb-twitter-GM2-ALL-ALL-150

| GM5 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 5 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 3.08 | 0.04 | 3.13 |
| | 1 | 6.07 | 22.95 | 91.79 | 91.79 | 0.00 | 0.03 | 212.63 |
| | 2 | 12.15 | 45.91 | 183.63 | 183.63 | 0.00 | 0.03 | 425.36 |
| | 4 | 24.31 | 91.84 | 367.35 | 367.35 | 0.00 | 0.03 | 850.88 |
| | 8 | 48.70 | 184.00 | 735.99 | 735.99 | 0.00 | 0.03 | 1704.71 |
| | 16 | 97.43 | 368.11 | 1472.43 | 1472.43 | 0.00 | 0.03 | 3410.44 |

Table P.13: nb-twitter-GM5-ALL-ALL-5

| GM5 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 10 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 3.08 | 0.09 | 3.17 |
| | 1 | 6.07 | 22.95 | 91.79 | 91.79 | 0.00 | 0.07 | 212.66 |
| | 2 | 12.15 | 45.91 | 183.63 | 183.63 | 0.00 | 0.07 | 425.40 |
| | 4 | 24.31 | 91.84 | 367.35 | 367.35 | 0.00 | 0.07 | 850.91 |
| | 8 | 48.70 | 184.00 | 735.99 | 735.99 | 0.00 | 0.07 | 1704.74 |
| | 16 | 97.43 | 368.11 | 1472.43 | 1472.43 | 0.00 | 0.07 | 3410.47 |

Table P.14: nb-twitter-GM5-ALL-ALL-10

| GM5 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 25 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 3.08 | 0.22 | 3.31 |
| | 1 | 6.07 | 22.95 | 91.79 | 91.79 | 0.00 | 0.17 | 212.76 |
| | 2 | 12.15 | 45.91 | 183.63 | 183.63 | 0.00 | 0.17 | 425.50 |
| | 4 | 24.31 | 91.84 | 367.35 | 367.35 | 0.00 | 0.17 | 851.01 |
| | 8 | 48.70 | 184.00 | 735.99 | 735.99 | 0.00 | 0.17 | 1704.85 |
| | 16 | 97.43 | 368.11 | 1472.43 | 1472.43 | 0.00 | 0.17 | 3410.57 |

Table P.15: nb-twitter-GM5-ALL-ALL-25

| GM5 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 50 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 3.08 | 0.45 | 3.53 |
| | 1 | 6.07 | 22.95 | 91.79 | 91.79 | 0.00 | 0.34 | 212.93 |
| | 2 | 12.15 | 45.91 | 183.63 | 183.63 | 0.00 | 0.34 | 425.67 |
| | 4 | 24.31 | 91.84 | 367.35 | 367.35 | 0.00 | 0.34 | 851.18 |
| | 8 | 48.70 | 184.00 | 735.99 | 735.99 | 0.00 | 0.34 | 1705.02 |
| | 16 | 97.43 | 368.11 | 1472.43 | 1472.43 | 0.00 | 0.34 | 3410.75 |

Table P.16: nb-twitter-GM5-ALL-ALL-50

| GM5 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 75 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 3.08 | 0.67 | 3.75 |
| | 1 | 6.07 | 22.95 | 91.79 | 91.79 | 0.00 | 0.51 | 213.10 |
| | 2 | 12.15 | 45.91 | 183.63 | 183.63 | 0.00 | 0.51 | 425.84 |
| | 4 | 24.31 | 91.84 | 367.35 | 367.35 | 0.00 | 0.51 | 851.35 |
| | 8 | 48.70 | 184.00 | 735.99 | 735.99 | 0.00 | 0.51 | 1705.18 |
| | 16 | 97.43 | 368.11 | 1472.43 | 1472.43 | 0.00 | 0.51 | 3410.91 |

Table P.17: nb-twitter-GM5-ALL-ALL-75

| GM5 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 150 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 3.08 | 1.34 | 4.42 |
| | 1 | 6.07 | 22.95 | 91.79 | 91.79 | 0.00 | 1.02 | 213.61 |
| | 2 | 12.15 | 45.91 | 183.63 | 183.63 | 0.00 | 1.02 | 426.35 |
| | 4 | 24.31 | 91.84 | 367.35 | 367.35 | 0.00 | 1.02 | 851.86 |
| | 8 | 48.70 | 184.00 | 735.99 | 735.99 | 0.00 | 1.02 | 1705.70 |
| | 16 | 97.43 | 368.11 | 1472.43 | 1472.43 | 0.00 | 1.01 | 3411.42 |

Table P.18: nb-twitter-GM5-ALL-ALL-150

| GB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 5 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 3.18 | 0.13 | 3.31 |
| | 1 | 3.21 | 12.11 | 48.44 | 48.44 | 0.00 | 0.21 | 112.40 |
| | 2 | 6.41 | 24.22 | 96.88 | 96.88 | 0.00 | 0.20 | 224.60 |
| | 4 | 12.82 | 48.44 | 193.78 | 193.78 | 0.00 | 0.20 | 449.03 |
| | 8 | 25.64 | 96.89 | 387.55 | 387.55 | 0.00 | 0.21 | 897.84 |
| | 16 | 51.31 | 193.85 | 775.39 | 775.39 | 0.00 | 0.20 | 1796.14 |

Table P.19: nb-twitter-GB3-ALL-ALL-5

| GB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 10 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 3.18 | 0.26 | 3.44 |
| | 1 | 3.21 | 12.11 | 48.44 | 48.44 | 0.00 | 0.41 | 112.61 |
| | 2 | 6.41 | 24.22 | 96.88 | 96.88 | 0.00 | 0.41 | 224.81 |
| | 4 | 12.82 | 48.44 | 193.78 | 193.78 | 0.00 | 0.41 | 449.24 |
| | 8 | 25.64 | 96.89 | 387.55 | 387.55 | 0.00 | 0.41 | 898.05 |
| | 16 | 51.31 | 193.85 | 775.39 | 775.39 | 0.00 | 0.41 | 1796.35 |

Table P.20: nb-twitter-GB3-ALL-ALL-10

| GB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Size (MB) | | | | | | |
| | | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 25 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 3.18 | 0.64 | 3.82 |
| | 1 | 3.21 | 12.11 | 48.44 | 48.44 | 0.00 | 1.02 | 113.22 |
| | 2 | 6.41 | 24.22 | 96.88 | 96.88 | 0.00 | 1.03 | 225.42 |
| | 4 | 12.82 | 48.44 | 193.78 | 193.78 | 0.00 | 1.03 | 449.86 |
| | 8 | 25.64 | 96.89 | 387.55 | 387.55 | 0.00 | 1.03 | 898.67 |
| | 16 | 51.31 | 193.85 | 775.39 | 775.39 | 0.00 | 1.03 | 1796.96 |

Table P.21: nb-twitter-GB3-ALL-ALL-25

| GB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Size (MB) | | | | | | |
| | | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 50 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 3.18 | 1.28 | 4.46 |
| | 1 | 3.21 | 12.11 | 48.44 | 48.44 | 0.00 | 2.05 | 114.24 |
| | 2 | 6.41 | 24.22 | 96.88 | 96.88 | 0.00 | 2.06 | 226.45 |
| | 4 | 12.82 | 48.44 | 193.78 | 193.78 | 0.00 | 2.06 | 450.88 |
| | 8 | 25.64 | 96.89 | 387.55 | 387.55 | 0.00 | 2.06 | 899.70 |
| | 16 | 51.31 | 193.85 | 775.39 | 775.39 | 0.00 | 2.06 | 1797.99 |

Table P.22: nb-twitter-GB3-ALL-ALL-50

| GB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Size (MB) | | | | | | |
| | | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 75 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 3.18 | 1.92 | 5.10 |
| | 1 | 3.21 | 12.11 | 48.44 | 48.44 | 0.00 | 3.08 | 115.28 |
| | 2 | 6.41 | 24.22 | 96.88 | 96.88 | 0.00 | 3.08 | 227.47 |
| | 4 | 12.82 | 48.44 | 193.78 | 193.78 | 0.00 | 3.08 | 451.91 |
| | 8 | 25.64 | 96.89 | 387.55 | 387.55 | 0.00 | 3.08 | 900.72 |
| | 16 | 51.31 | 193.85 | 775.39 | 775.39 | 0.00 | 3.09 | 1799.02 |

Table P.23: nb-twitter-GB3-ALL-ALL-75

| GB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group Size | Web1T % | Size (MB) | | | | | | |
| | | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 150 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 3.18 | 3.84 | 7.02 |
| | 1 | 3.21 | 12.11 | 48.44 | 48.44 | 0.00 | 6.17 | 118.37 |
| | 2 | 6.41 | 24.22 | 96.88 | 96.88 | 0.00 | 6.18 | 230.57 |
| | 4 | 12.82 | 48.44 | 193.78 | 193.78 | 0.00 | 6.19 | 455.02 |
| | 8 | 25.64 | 96.89 | 387.55 | 387.55 | 0.00 | 6.18 | 903.82 |
| | 16 | 51.31 | 193.85 | 775.39 | 775.39 | 0.00 | 6.17 | 1802.10 |

Table P.24: nb-twitter-GB3-ALL-ALL-150

| OSB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 5 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 7.30 | 0.26 | 7.56 |
| | 1 | 12.41 | 46.89 | 187.54 | 187.54 | 0.00 | 0.41 | 434.80 |
| | 2 | 24.82 | 93.77 | 375.10 | 375.10 | 0.00 | 0.41 | 869.20 |
| | 4 | 49.65 | 187.57 | 750.28 | 750.28 | 0.00 | 0.41 | 1738.19 |
| | 8 | 99.29 | 375.14 | 1500.55 | 1500.55 | 0.00 | 0.41 | 3475.94 |
| | 16 | 198.81 | 751.03 | 3004.14 | 3004.14 | 0.00 | 0.42 | 6958.53 |

Table P.25: nb-twitter-OSB3-ALL-ALL-5

| OSB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 10 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 7.30 | 0.52 | 7.83 |
| | 1 | 12.41 | 46.89 | 187.54 | 187.54 | 0.00 | 0.83 | 435.22 |
| | 2 | 24.82 | 93.77 | 375.10 | 375.10 | 0.00 | 0.83 | 869.62 |
| | 4 | 49.65 | 187.57 | 750.28 | 750.28 | 0.00 | 0.83 | 1738.61 |
| | 8 | 99.29 | 375.14 | 1500.55 | 1500.55 | 0.00 | 0.83 | 3476.36 |
| | 16 | 198.81 | 751.03 | 3004.14 | 3004.14 | 0.00 | 0.83 | 6958.95 |

Table P.26: nb-twitter-OSB3-ALL-ALL-10

| OSB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 25 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 7.30 | 1.31 | 8.61 |
| | 1 | 12.41 | 46.89 | 187.54 | 187.54 | 0.00 | 2.09 | 436.47 |
| | 2 | 24.82 | 93.77 | 375.10 | 375.10 | 0.00 | 2.08 | 870.87 |
| | 4 | 49.65 | 187.57 | 750.28 | 750.28 | 0.00 | 2.08 | 1739.86 |
| | 8 | 99.29 | 375.14 | 1500.55 | 1500.55 | 0.00 | 2.08 | 3477.61 |
| | 16 | 198.81 | 751.03 | 3004.14 | 3004.14 | 0.00 | 2.09 | 6960.20 |

Table P.27: nb-twitter-OSB3-ALL-ALL-25

| OSB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 50 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 7.30 | 2.60 | 9.90 |
| | 1 | 12.41 | 46.89 | 187.54 | 187.54 | 0.00 | 4.17 | 438.56 |
| | 2 | 24.82 | 93.77 | 375.10 | 375.10 | 0.00 | 4.18 | 872.97 |
| | 4 | 49.65 | 187.57 | 750.28 | 750.28 | 0.00 | 4.16 | 1741.94 |
| | 8 | 99.29 | 375.14 | 1500.55 | 1500.55 | 0.00 | 4.16 | 3479.69 |
| | 16 | 198.81 | 751.03 | 3004.14 | 3004.14 | 0.00 | 4.17 | 6962.28 |

Table P.28: nb-twitter-OSB3-ALL-ALL-50

| OSB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 75 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 7.30 | 3.90 | 11.20 |
| | 1 | 12.41 | 46.89 | 187.54 | 187.54 | 0.00 | 6.28 | 440.66 |
| | 2 | 24.82 | 93.77 | 375.10 | 375.10 | 0.00 | 6.22 | 875.01 |
| | 4 | 49.65 | 187.57 | 750.28 | 750.28 | 0.00 | 6.27 | 1744.05 |
| | 8 | 99.29 | 375.14 | 1500.55 | 1500.55 | 0.00 | 6.26 | 3481.79 |
| | 16 | 198.81 | 751.03 | 3004.14 | 3004.14 | 0.00 | 6.25 | 6964.37 |

Table P.29: nb-twitter-OSB3-ALL-ALL-75

| OSB3 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Size (MB) | | | | | | |
| Group Size | Web1T % | keys.mph | signature | counts | logprobs | vocabmap | Authors Model | TOTAL |
| 150 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 7.30 | 7.79 | 15.10 |
| | 1 | 12.41 | 46.89 | 187.54 | 187.54 | 0.00 | 12.51 | 446.89 |
| | 2 | 24.82 | 93.77 | 375.10 | 375.10 | 0.00 | 12.47 | 881.26 |
| | 4 | 49.65 | 187.57 | 750.28 | 750.28 | 0.00 | 12.46 | 1750.24 |
| | 8 | 99.29 | 375.14 | 1500.55 | 1500.55 | 0.00 | 12.50 | 3488.02 |
| | 16 | 198.81 | 751.03 | 3004.14 | 3004.14 | 0.00 | 12.47 | 6970.58 |

Table P.30: nb-twitter-OSB3-ALL-ALL-150

THIS PAGE INTENTIONALLY LEFT BLANK

# Referenced Authors

THIS PAGE INTENTIONALLY LEFT BLANK

# Initial Distribution List

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudly Knox Library
   Naval Postgraduate School
   Monterey, California