Data Sharing and Resampled LASSO: A word based sentiment Analysis for IMDb data

Ashutosh K. Maurya

Applied Statistics Unit
Indian Statistical Institute
Kolkata 700 108
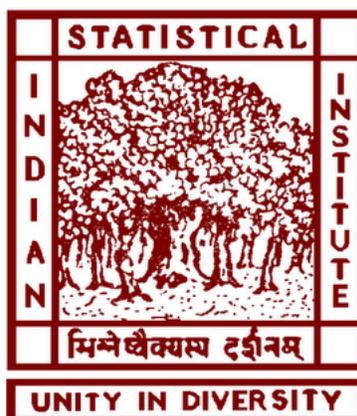
DATA SHARING AND RESAMPLED LASSO: A WORD BASED SENTIMENT ANALYSIS FOR IMDB DATA

ASHUTOSH K. MAURYA[1,2]

APPLIED STATISTICS UNIT,

INDIAN STATISTICAL INSTITUTE, KOLKATA

ABSTRACT. In this article we study variable selection problem using LASSO with new improvisations. LASSO uses ℓ_1 penalty, it shrinks most of the coefficients to zero when number of explanatory variables (p) are much larger the number of observations (N). Novelty of the approach developed in this article blends basic ideas behind resampling and LASSO together which provides a significant variable reduction and improved prediction accuracy in terms of mean squared error in the test sample. Different weighting schemes have been explored using *Bootstrapped LASSO*, the basic methodology developed in here. Weighting schemes determine to what extent of data blending in case of grouped data. Data sharing (DSL) technique developed by [11] lies at the root of the present methodology. We apply the technique to analyze the IMDb dataset as discussed in [11] and compare our result with [11].

Key Words : High dimensional regression, Sentiment analysis, Weighted LASSO, Resampling, Cross validation, Data sharing.

1. INTRODUCTION

Extracting important features from any big data is a challenging task. There are number of algorithms for variable selection. In [3] a notion of data enriched regression has been developed which along with other variable selection techniques (e.g. Ridge [13], Subset selection [2], Least Angle Regression(LARS) [6], Least Absolute Shrinkage and Selection Operator(LASSO) [22] , De-noising [5]) which give rise to a host of statistical algorithms for feature selection in big data. For the purpose of this article we study different exploratory variations of regression LASSO based feature selection as developed by [11]. We introduce two major variations to the DSL algorithm. The first one uses resampling (bootstrap) which provides us with a measure of variation when we use the algorithm by resampling the training sample. The second variation is about choice of weighting scheme across different groups in regularization.

The idea of resampling has a rich history ([15], [20], [21], [23], [7]). Traditionally it comes in two forms: Jackknife and Bootstrap. "The idea of jackknife resampling has evolved into cross validation technique which is commonly used in lasso regularization." The idea of bootstrap, its variations and applications in different context can be found in ([7], [8], [17]).

The IMDb problem can be viewed as an application of sentiment analysis where the goal is to extract a smaller set of words from very large *bag-of-words* that have better classification of movie ratings. Following [11] we consider only linear classifiers so that it becomes a linear regression with sparse data. In this study rating have been converted into binary (positive and negative sentiments). This template example can serve as a prototype of a large number of similar problems of sentiment analysis in different areas. The IMDb data used in our analysis consist of 50k movie reviews on a scale of allowing no more than 30 reviews per movie that have been divided into training and test sets of equal size. In our data analysis, we follow the similar resampling procedure to create training and test data sets as outlined in [11]. It is noteworthy that one can use resampling procedure on the entire datasets, however this will lead to higher computational complexity.

Related literature includes works from Opinion mining, Natural Language Processing, Machine learning among others. Probabilistic classifiers for tracking point of view was designed by [24]. A machine learning approach to predict the semantic orientation of adjectives (positive or negative) was developed by [12]. Substantial research has been done on sentiment analysis. It has been done over various topics such as sentiment analysis over movie reviews [19], product reviews ([4], [18]), news and blogs ([10], [1]). Following [19] conducted an extensive experiment on movie reviews using three traditional supervised machine learning methods (i.e., Naive Bayes (NB), maximum entropy classification (ME), and support vector machines (SVM)). Following [19], results indicate that standard machine learning techniques definitively outperform human-produced baselines. However, [19] found that machine learning methods could not perform as well on sentiment classification as on traditional topic-based categorization.

The article is presented as follows: In Section (2) we present basic theoretical preliminaries about variable selection of high dimensional data. In this section we shall discuss LASSO and other related regularization methods used for this purpose and how we fit our proposed resampling scheme in this framework. In section (3) we illustrate our proposed approaches and methodology for the analysis of the data coming from more than two different sources. We then in section (4) talk about the preparation of the dataset for analysis through an example to apply our methodology. Related experimental results are discussed in section (5) and finally some discussion is given in Section (6).

2. PRELIMINARIES

The basic setup starts with a multiple linear regression model for a set of ratings $y_1, y_2, ..., y_n$ with associated explanatory variables $x_1, x_2, ..., x_n$. The explanatory variables are high dimensional that is, $x_i = (x_{i1}, x_{i2}, ..., x_{ip})$ where $p >> n$. Moreover we focus on the situation where the design matrix $X = ((x_{ij}))$ is binary and sparse.

We consider the multiple linear regression model of y_i on x_i with intercept. Also while the rating scale is integral valued and bounded which treat y_i's as real numbers.

$$y_i = \mu + x_i^T \beta + \epsilon_i,$$

ϵ_i's are independently distributed with mean zero and variance σ^2 for $1 \leq i \leq n$. and parameter vector $\beta = (\beta_1, \beta_2, ..., \beta_p)$.

Furthermore if the data is stratified into $G \geq 2$ different groups we assign a group indicator variable g_i for each observation and thus the regression sum of squares (SS) can be written in the following way

$$\sum_{i=1}^{n} (y_i - x_i^T(\beta + \Delta_{g_i}))^2.$$

Following [11] the coefficients $\Delta = (\Delta_1, \Delta_2, \ldots, \Delta_g)$ represents group effects.

It is well known that ordinary least square (OLS) estimation procedure performs poorly in both prediction and interpretation when $(p >> n)$. The unknown intercept and parameter vector can be estimated reasonably well, if β is sparse in some sense. The sparsity can be quantified in terms of the ℓ_q norm for $1 \leq q \leq \infty$. The ℓ_0-analogue $\|\beta\|_0^0 = |\{i; \beta_i \neq 0\}|$ (which is not a norm) which counts the number of non-zero entries of the parameter. The notation $\|\beta\|_0^0$ (where $0^0 = 0$)is in analogy to $\|\beta\|_q^q = \sum_{i=1}^{p} |\beta_i|^q$ for $0 < q < \infty$.

To overcome the limitation of OLS in high dimensional setting, we include a regularization term that penalizes the model over coefficients. Subset selection, ridge regression, LASSO and elastic net are important techniques for improving OLS estimates. Subset regression [17] provides interpretable model. Since it is a discrete process so either variables are retained or dropped form the model hence the results can be extremely variable. Commonly used techniques are based on least squares and a penalty which involves the number of parameters in the candidate sub-model;

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \sum_i (y_i - x_i^T \beta)^2 + \lambda \|\beta\|_0^0,$$

where the ℓ_0-penalty is $\|\beta\|_0^0 = \sum_{i=1}^{p} 1(\beta_i \neq 0)$. The regularization parameter λ is usually chosen via some cross-validation scheme. Under ℓ_0-penalty the estimator is infeasible to compute when p is of large size since the ℓ_0-penalty is a non-convex function of β [2]. Many other well known model selection criteria such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) fall into this framework.

In another approach, [13] have proposed a regression model which minimizes the residual sum of squares subject to a bound on the L_2- norm of the coefficient. As ridge regression is continuous shrinkage method, it achieves better predictive performance for the case $(p >> n)$. But there is a drawback with ridge regression, it does not set any coefficient to zero and hence does not give any interpretable model [22]. To overcome with this drawback [22] proposed a new technique for variable selection when $(p >> n)$ called LASSO. The LASSO uses l_1 penalty which is convex surrogate of l0-norm, results in convex optimization problem and a number of fast algorithms can be used to solve this problem. LASSO shrinks small coefficients to zero and thus performs model selection. In doing so, it results in some bias in the estimated coefficients yet often improves the predictive performance of the model.

The LASSO does poorly for the case $(p >> n)$ or if there are high correlation between predictors [25]. To deal with this problem [25] suggested (elastic net) a new regularization and variable selection method which is convex linear combination of ℓ_1 and ℓ_2 norm. The elastic net enjoys property of both ridge and LASSO regression. This regression technique is much useful in the case when $(p >> n)$. Elastic net solves the following problem

$$\hat{\beta}_{en} = \arg\min_{\beta} \sum_i (y_i - x_i^T \beta)^2 + \lambda P_\alpha(\beta),$$

where
$$P_\alpha(\beta) = (1 - \alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1,$$
is the elastic net penalty [25]. Which is equivalent to LASSO when $(\alpha = 1)$ and equivalent to ridge regression when $(\alpha = 0)$ and for all α i.e.$\in [0, 1]$.

Following [3] proposed a new regularization technique for prediction when the data set come from two different sources. In [3], the authors proposes combining two linear regressions on two different datasets, one dataset is used for estimating the regression coefficients, and the squared error penalty term for second dataset is used to improve the estimates of regression coefficients. The data enriched regression solves the following problem

$$(\hat{\beta}, \hat{\Delta}) = \arg\min_{(\beta,\Delta)} \sum_{i:g_i=1} (y_i - x_i^T \beta)^2 + \sum_{i:g_i\neq 1} (y_i - x_i^T(\beta + \Delta))^2 + \lambda P(\Delta),$$

where Δ is same size as β and

$$P(\Delta) = \sum_{i:g_i=1} (x_i^T \Delta)^2,$$

is the quadratic penalty. Here regularization is done over Δ only. The above model is not useful to study over different groups together.

In [11] proposed a new model which is generalization of data enriched regression. The DSL can be used for multiple groups, which are coming from different non-overlapping pre-specified groups. In DSL, the ℓ_1-penalty over both β and Δ_g encourages the sparsity and improved estimation in β simultaneously. Suppose we have n observation coming from G groups, and let p be the number of predictors in each group then the generalized data enriched model can be written as

(2.1) $$y_i = x_i^T(\beta + \Delta_{g_i}) + \epsilon_i,$$

where ϵ is the error term term independent of x and β.

The DSL minimizes the following

(2.2) $$(\hat{\beta}, \hat{\Delta}_1, ..., \hat{\Delta}_g) = \arg\min \frac{1}{2} \sum_i (y_i - x_i^T(\beta + \Delta_{g_i}))^2 + \lambda \left(\|\beta\|_1 + \sum_{g=1}^{G} r_g\|\Delta_g\|_1 \right),$$

where λ is the usual regularization parameter. Here r_g is used as regularization parameter over groups which controls the amount of sharing between the groups.

2.1. De-noising. To make more precise the interpretation of the extracted features another idea (by removing the noise through regularization) is proposed in this paper. The idea of De-noising was proposed by [5]. Following [5] suggested thresholding procedure for recovering functions for noisy data. They applied soft thresholding nonlinearity $\eta_y(t) = sgn(y)(|y| - t)_+$ coordinate-wise to the empirical wavelet coefficient. In [5] the chosen threshold is

(2.3) $$t_n = \sqrt{2log(n)}.\gamma_1.\sigma/\sqrt{n},$$

where γ_1 is a constant which is defined in [5].

2.2. Differential Regularization. Choice of regularization parameter is a challenging problem in high dimensional data analysis. Regularization has clear benefit in producing sparse solution as well reduces false discovery rate. A larger value of r_g accounts for large amount of sharing between the groups and a smaller value of r_g accounts there will be less amount of

sharing between the groups. Following [11] suggested some values of r_g that is if $\sum_g r_g < 1$ it will be equivalent to separate regression, If $\sum_g r_g = 1$ then there will be identifiability concern, and a default value of r_g which is $1/\sqrt{G}$ when there are same number of observation in each group. From the above conditions it is clear that $\sum_g r_g$ has to be more than one for data shared lasso. In this article we are proposing an idea for choosing the value of r_g which results gain in mean squared error than [11].

3. Methodology

We propose two type of resampling techniques to reduce the size of the dataset ensuring that it conveys similar information concisely. Proposed technique takes care of multicollinearity, removes redundant features, fastens the time required for performing same computations and it is helpful in noise removal also and as result of that we can improve the performance of models. In addition, we propose new methods of selecting regularization parameters. The data analysis shows the empirical evidence of significance of proposed resampling technique and new method of selecting regularization parameter. We keep the methodology adaptive by taking care of unequal sample sizes in different groups. This will be discussed in greater detail when we analyze the experimental results.

3.1. Resampling Technique.
In this article we propose two types of resampling schemes, Bootstrapped LASSO Scheme (BLS) and Bootstrapped Shared LASSO Scheme (BSLS) or Bootstrapped DSL. Here we consider the estimation accuracy of the parameter β, a different task than prediction. Under compatibility assumptions on the design matrix X and on the sparsity $s_{g_i,0} = |S_{g_i,0}|$ in a linear model in G groups, let $S_{g_i,0}$ denote the number of active set of variables initially, defined as

$$(3.1) \qquad S_{g_i,0}(\lambda_{g_i}) = \{j;\ \beta_{g_i,j}^0 \neq 0.\ j = 1, \ldots, p\},$$

and the set of estimated variables using LASSO be given as

$$(3.2) \qquad \hat{S}_{g_i}(\lambda_{g_i}) = \{j;\ \hat{\beta}_{g_i,j}^0(\lambda_{g_i}) \neq 0.\ j = 1, \ldots, p\}.$$

For data analysis, we created training $(T_{r_{g_i}})$ and test $(T_{s_{g_i}})$ sets of equal size via simple random sampling. The train set $(T_{r_{g_i}})$ is taken under study and we separated it into G groups $(X_1, X_2, \ldots, X_G)$. With replacement samples $(X_1^*, X_2^*, \ldots, X_G^*)$ are taken from each group X_G, where $X_{g_i}^*$ contains hundred with replacement samples $\left(\left((x_{g_i,1}^*)\right), \left((x_{g_i,2}^*)\right), \ldots, \left((x_{g_i,100}^*)\right) \right)$. As we know the LASSO is a convex optimization problem which uses ℓ_1-penalty and solves the following problem

$$(3.3) \qquad \hat{\beta}_{lasso}(\lambda) = \arg\min_{\beta} \sum_i (y_i - x_i^T \beta)^2 + \lambda \|\beta\|_1,$$

where λ is usual regularization parameter which is chosen via 10 fold-CV. We then implemented the LASSO algorithm (3.3) over the bootstrapped samples $\left(\left(x_{g_i,k}^*\right) \right)$ where $1 \leq k \leq 100$. This gives us hundred different set of active set of variables $\hat{S}_{g_i,k}^*(\lambda_{g_i,k}^*)$. The union of the set $\hat{S}_{g_i,k}^*(\lambda_{g_i,k}^*)$ resulted in a lower dimensional dataset $X_{g_i}^{**}$. The new designed matrix

$X_{g_i}^{**}$ can be expressed as;

$$(3.4) \qquad X_{g_i}^{**} = \bigcup_{k=1}^{100} S_{g_i,k}^*(\lambda_{g_i,k}^*).$$

The above procedure is repeated for other groups over the bootstrapped sample $(X_1^*, X_2^*, \ldots, X_G^*)$ and we get different sets of active set variables that results in different lower dimensional datasets $(X_1^{**}, X_2^{**}, \ldots, X_G^{**})$.

In addition to this approach we performed another bootstrapped resampling over the dataset X called bootstrapped DSL. For our purpose primarily we divide the dataset into two halves, and then we separate the dataset group wise and take with replacement samples from each group $(X_1^*, X_2^*, \ldots, X_G^*)$ each of size hundred from the train set T_{rgi}. Where

$$X_1^* = \Big(\big(\, (x_{1,1}^{**}) \, \big), \big(\, (x_{1,2}^{**}) \, \big), \ldots, \big(\, (x_{1,100}^{**}) \, \big) \Big),$$
$$X_2^* = \Big(\big(\, (x_{2,1}^{**}) \, \big), \big(\, (x_{2,2}^{**}) \, \big), \ldots, \big(\, (x_{2,100}^{**}) \, \big) \Big),$$
$$\ldots,$$
$$X_G^* = \Big(\big(\, (x_{G,1}^{**}) \, \big), \big(\, (x_{G,2}^{**}) \, \big), \ldots, \big(\, (x_{G,100}^{**}) \, \big) \Big).$$

To implement the DSL algorithm we prepared the data matrices Z_k^* (where $1 \leq k \leq 100$) as defined in [11], with the help of the bootstrapped samples $(X_1^*, X_2^*, \ldots, X_G^*)$, as $(Z_1^*, Z_2^*, \ldots, Z_{100}^*)$. The DSL algorithm is implemented over each bootstrapped sample (Z_k^*) which gives us different set of active set variables $\hat{S}_z^*(\lambda^*) = \Big(\hat{S}_{z_1}^*(\lambda_1^*), \hat{S}_{z_2}^*(\lambda_2^*), \ldots, \hat{S}_{z_{100}}^*(\lambda_{100}^*) \Big)$. The union of the set $\hat{S}_z^*(\lambda^*)$, resulted in a dataset Z^{**} having lower dimension which can be expressed as

$$(3.5) \qquad Z^{**} = \bigcup_{k=1}^{100} \hat{S}_{z_i}^*(\lambda_{z_i}^*),$$

to check the significance of proposed resampling procedures, we implement LASSO and DSL algorithms on the resulting data sets. The results are explained in section (5).

3.2. **Improvisation in Data Shared Lasso Model.** Choice of the regularization parameter r_g in DSL model plays an important role in the sense that it controls the amount of sharing within the groups. The DSL model can be seen as LASSO model [11], with the following representation:

$$(3.6) \qquad \frac{1}{2}\|\tilde{y} - Z\tilde{\beta}\|^2 + \lambda\|\tilde{\beta}\|_1 = \frac{1}{2}\sum_i \big(y_i - x_i^T(\beta + \Delta_{g_i}) \big)^2 + \lambda \left(\|\beta\|_1 + \sum_{g=1}^{G} r_g\|\Delta_g\|_1 \right).$$

The above representation allows to apply LASSO algorithm directly to solve the DSL optimization problem. We apply de-noising threshold on the selected LASSO model from equation (3.6). By varying the value of $\gamma_1 \in [0, 1/2]$ in (2.3), we observed the variability in the mean squared error. A pictorial representation of the variability in MSE by varying the constant γ_1 is shown through figure (1).

From the following figure (1) we see that by varying the value of $\gamma_1 \in [0, 1/2]$, after a certain point the MSE is increasing rapidly and the minimum of the mean squared error is
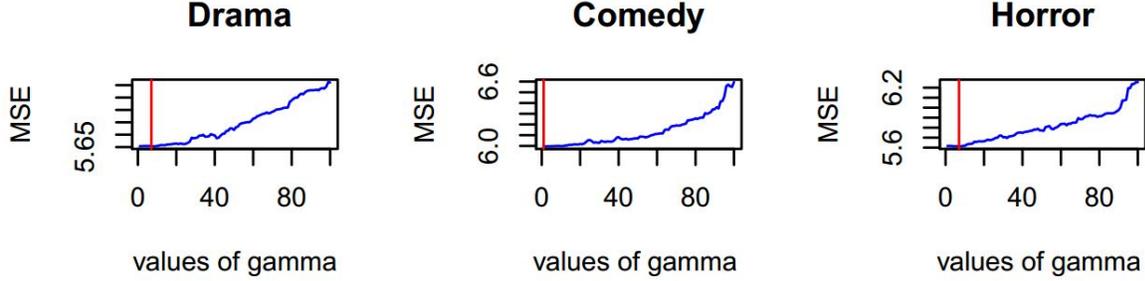
FIGURE 1. These plots are showing the variability in the mean squared by varying the value of γ_1 in Donoho's global de-noising (2.3) threshold. Here x-axis is showing the values of γ_1, a vector of size hundred ($\gamma_1 \in [0, 1/2]$) and the relative change in MSE is in y-axis. The red vertical line is showing the point at which the minimum of the mean squared error occurred.

attained at a different point other than ($\gamma_1 = 0$), which indicates the presence of noise in the extracted features.

The objective function (2.2) is overparameterized as typically cardinality of β and Δ_g is larger than sample size n. Therefore we need to penalize the Δ_g term by ℓ_1 penalty to reduce the effective number of parameters. Following [11], for $r_g = r$, the penalty term simplifies to,

$$\lambda(\|\beta\|_1 + \sum_g r_g \|\Delta_g\|_1) = \lambda(\|\beta\|_1 + \sum_g r\|\Delta_g\|_1)$$

$$= \lambda r(\|\beta\|_1/r + \sum_g \|\Delta_g\|_1 = \lambda^{new}(\|\beta\|_1/r + \sum_g \|\Delta_g\|_1)$$

where $\lambda^{new} = \lambda r$.

Let us consider the case when we have unequal number of sample sizes i.e. $n_{g_1} \neq n_{g_2} \neq \cdots \neq n_G$, where n_{g_i} denote the sample size of each group. Certain arguments have been given in [11] for choosing the value of regularization parameter r_g.

Following [11] has shown that if we assume that we have fixed $\beta_g^* \in \mathcal{R}^p$ and r_g such that $\beta_g^* = \beta + \Delta_g$ then the problem

(3.7) $$(\hat{\beta}, \hat{\Delta}_g) = \lambda(\|\beta\|_1 + \sum_g r_g\|\Delta_g\|_1),$$

is separable in p, for the case $p = 1$, all values that satisfy the equality constraint fall into a family that can be characterized as $\beta = c$, and $\Delta_g = \beta_g^* - c$ hence (3.7) is equivalent to solving unconstrained optimization in one variable: $\hat{c} = \arg\min |c| + \sum_g r_g|\beta_g^* - c|$ and the constraint $\sum_g r_g < 1$ leads to separate regression. From the above discussion it is clear that the value of r_g would be chosen under the condition $\sum_g r_g > 1$. When we have unequal number of observations the above problem (3.7) can be simplified as:

$$\lambda(\|\beta\|_1 + \sum_g r_g\|\Delta_g\|_1) = \lambda(\|\beta\|_1 + r_1\|\Delta_1\|_1 + r_2\|\Delta_2\|_1, \cdots, r_G\|\Delta_G\|_1)$$
$$= \lambda\|\beta\|_1 + \lambda(r_1\|\Delta_1\|_1 + r_2\|\Delta_2\|_1, \cdots, r_G\|\Delta_G\|_1)$$
$$= \lambda\|\beta\|_1 + r_1^*\|\Delta_1\|_1 + r_2^*\|\Delta_2\|_1, \cdots, r_G^*\|\Delta_G\|_1,$$

8

where $r_{g_i}^* = \lambda r_{g_i}$ are new regularization parameter over the groups. As we have unequal number of observation so the chosen penalty should depend on the sample size under the condition $\sum_g r_g > 1$. Suggested regularization parameters are chosen with the help of universal global de-noising threshold (2.3) that reduces the noise from the extracted features. The penalties used in the DSL model are following $\sqrt{\frac{1}{3}}$, $\sqrt{\frac{n_{g_i}}{N}}$, $\sqrt{\frac{\log N}{\log n_i}}$, $\frac{N}{n_{gi}}$, $\frac{\log N}{\log n_{gi}}$, $\frac{\log n_{gi}}{\log N}$ and $\sqrt{\frac{\log n_{g_i} \times N}{\log N \times n_{g_i}}}$ respectively over different groups, where $N = n_{g_1} + n_{g_2} + n_{g_3}$.

3.3. **A Combined Study of DSL and Separate LASSO Together With Differential Regularization Scheme.** Another attempt to reduce the noise from the extracted features by using both DSL and LASSO together, proposed in this article. For our purpose firstly we implement LASSO algorithm over groups $(X_1, X_2, \ldots, X_G)$ separately which gives us different set of active set variables $\hat{S}_{g_i}^*(\lambda_{g_i}^*)$ where $1 \leq g_i \leq G$. Next we implemented the DSL algorithm over the matrix Z together with differential regularization which gives us another different set of active set variables $\hat{S}_{z,(r_g)_j}^*(\lambda_{(r_g)_j}^*)$. Here we are interested in shared active set variables in the set $\hat{S}_{z,(r_g)_j}^*(\lambda_{(r_g)_j}^*)$ and $\hat{S}_{g_i}^*(\lambda_{g_i}^*)$ for $1 \leq g_i \leq G$. Let the set of shared active set variables be denoted as $\hat{S}_{z_s,(r_g)_j}^*(\lambda_{(r_g)_j}^*)$. The main idea is to create different subgroups by using both set of active set variables $\hat{S}_{g_i}^*(\lambda_{g_i}^*)$ and $\hat{S}_{z_s,(r_g)_j}^*(\lambda_{(r_g)_j}^*)$ and study the relative change in MSE by removing different subgroups from the main model. These subgroups are created by taking the intersection between different groups which are as follows
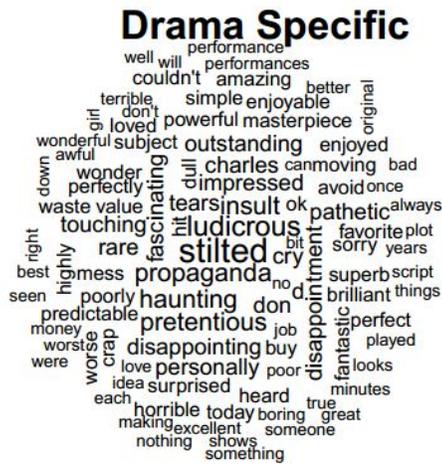
$$\left(\hat{S}_{g_i}^*(\lambda_{g_i}^*) \right) \bigcap \left(\hat{S}_{z_s,(r_g)_j}^*(\lambda_{(r_g)_j}^*) \right),$$
$$\left(\bigcap_{g_i=1}^{G} \hat{S}_{g_i}^*(\lambda_{g_i}^*) \right) \bigcap \left(\hat{S}_{z_s,(r_g)_j}^*(\lambda_{(r_g)_j}^*) \right),$$
$$\left(\hat{S}_{z_s,(r_g)_j}^*(\lambda_{(r_g)_j}^*) \right) - \left(\bigcup_{g_i}^{G} \hat{S}_{g_i}^*(\lambda_{g_i}^*) \right),$$

where $g_i(= 1, 2, ..., G)$ denote the number of groups and $j(= 1, 2, ..., l)$ denote the number of penalty used. The above proposed technique help us to classify all those set of features which improves the prediction accuracy.
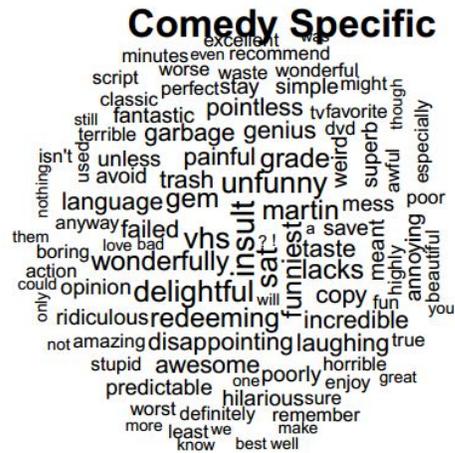
4. Experimental Setup for IMDb Data

For our purpose we used publicly available IMDb dataset review dataset aclImdb [14] of movie reviews from `IMDb.com`. which contains set of reviews and corresponding ratings associated with the them where the dependent variable Y_i's are integer rating and explanatory variable X_i's are text features. The text features are converted into integers by taking the number of occurrences of each feature in each review. This resulted in a high dimensional dataset that is the sample size is less than the features. Later the entries of the predictor matrix X are converted into binary matrix and the sparse representation of the data matrix X is used for the analysis that is

(4.1)
$$x_{ij} = \begin{cases} 1, & \text{if } j^{th} \text{ feature is present in } i^{th} \text{ review} \\ 0, & \text{otherwise.} \end{cases}$$

Drama Specific

(a) Drama

Comedy Specific

(b) Comedy

Horror Specific
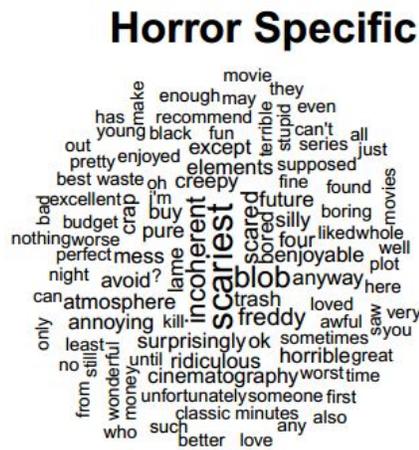
(c) Horror

FIGURE 2. This word cloud is for the coefficients found after running Boot-strapped LASSO model separately on each genre. The size of the words are indicating proportion of words appeared in 100 bootstrap sample

As there are 2^p possible sub-models, computational feasibility is crucial and when the data passes through the LASSO [22] it selects at most n variables before it saturates. To improve the computation feasibility of the model we need to compress the size of the bag of features.

Let us assume that X_G come from $N_p(0, \Sigma)$, where $\Sigma \succ 0$. We divide the dataset X_G into two halves, train $T_{r_{g_i}}$ and test $T_{s_{g_i}}$ set. The DSL algorithm is formulated for the supervised learning problem which is applied to a sentiment analysis dataset. This dataset contains 50k movie reviews, allowing no more than 30 reviews per movie that have been divided into two sets of equal size training and test sets. This dataset contains polarized reviews half of the reviews are positive (rating ≥ 7) and other half are negative (rating ≤ 4). Binary *bag-of-words* representation of the reviews is used. Only those words that appear in at least five reviews in our training set, are considered for the analysis of the dataset. This resulted $p = 27743$ features in our dataset. Response value is integer rating. To implement the DSL algorithm, we mainly focused on genres of the movies. The DSL model is fitted over three most popular genres drama, comedy and horror. Only those review are considered which appear in at least one of the three genres. This resulted in $n = 16386$ reviews in training set (8286 dramas, 5027 comedies, 3073 horror movies). The DSL algorithm can be implemented with any lasso solver using a straightforward augmented data approach:

Following [11] defined the data matrix Z as

$$
Z = \begin{bmatrix}
X_1 & r_1 X_1 & 0 & . & . & . & 0 \\
X_2 & 0 & r_2 X_2 & . & . & . & 0 \\
. & & & & & & \\
. & & & & & & \\
. & & & & & & \\
X_G & 0 & 0 & . & . & . & r_G X_G
\end{bmatrix},
$$

where X_j and y_j denote the data matrix for the group $g_i = j$. The above problem is solved by R package **glmnet** [9] which has capability to use sparse representation of a data matrix. It is quite faster than other LASSO solver. Here it is easy to change the family and loss function. In **glmnet** package we can fit ridge as well as elastic net by changing the values of α. For ($\alpha = 1$) glmnet fits LASSO model, for ($\alpha = 0$) it fits ridge regression and for $\alpha \in (0, 1)$ it fits elastic net model.

5. EXPERIMENTAL RESULT

We implemented our methodology to IMDb dataset, we performed the data analysis for the various choice of the group regularization parameter r_g.

Following fig.(3) shows the nonzero coefficients appeared in the main model. Blue vertical lines are showing the presence of nonzero coefficient in both the main model. First three pictures are showing for LASSO and fourth is showing for DSL model. Following fig.(3) showing the results obtained after fitting the Bootstrapped LASSO and Bootstrapped DSL. Here each horizontal red line is showing the number of times a coefficient appeared in hundred bootstrap sample. First three pictures in fig.(3) are obtained after running bootstrapped LASSO in each genre and the fourth picture is showing the results obtained after running bootstrapped DSL.
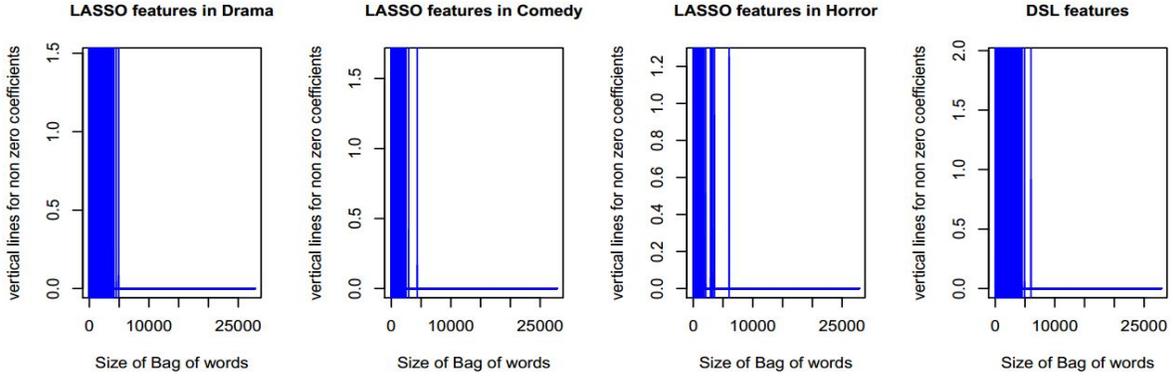
FIGURE 3. Here the vertical lines are showing the presence of the non zero coefficients after running LASSO in each genre and DSL on Z.
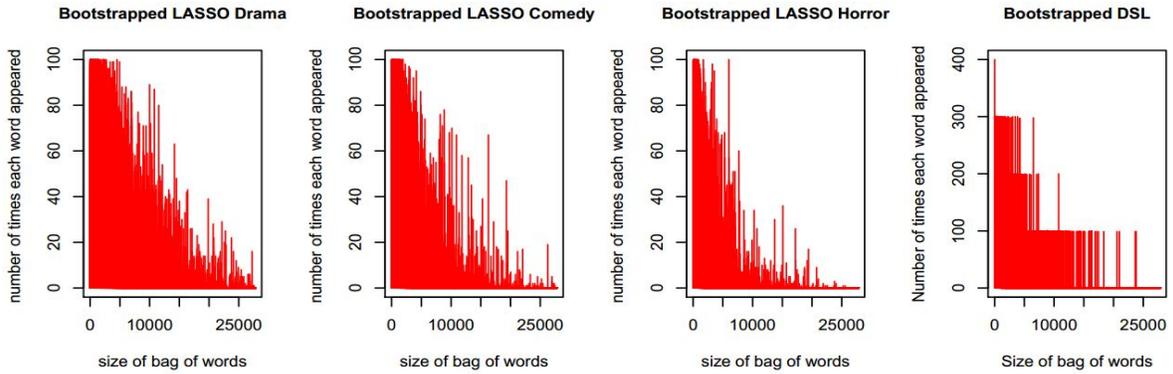

FIGURE 4. Here are the number of times each word found from running bootstrapped LASSO on each genre and bootstrapped DSL on matrix Z. Here x-axis is showing the size of the bag of words and y axis is showing number of times each word appeared.

From both the figure (3) and (4) we see that there are many features which appeared more than sixty times while they did not appear in the main model. This is the one of the advantage of our proposed technique that it gives us all the important sentiments which are in the dataset but not selected in the main model. These sentiments can be extracted by the proper choice of the threshold over the fraction of the extracted features. Another advantage of this technique is that the union of the all the extracted features resulted in a lower dimensional dataset hence it can be think of as a one of the dimension reduction technique. This reduces the size of the *bag-of-features* for drama it reduces to 9854 from 27743 for comedy it reduces to 6201 and for horror it reduces to 4114 features from 27743 features and also the column size of the entries of the matrix Z [11], also reduces to 11295 from 27743 features.

Proposed methodology gives us two types of dataset of different dimension. We performed data analysis over both the dataset the results are given in table (1) and (2). We run LASSO in each genre and DSL on Z on both type of dataset and results are shown in table (1) and (2).

Here we can see that the prediction accuracy of both the data set after running LASSO and DSL model is same as main dataset. These dataset fastens the time required for performing same data analysis and removes the redundant features.

TABLE 1. Comparison of test set mean squared error after running LASSO in each genre of both type of dataset with main dataset.

	Drama	Comedy	Horror
Old data	5.66	5.99	5.63
BLS	5.66	5.99	5.63
BSLS	5.66	5.99	5.63

TABLE 2. Comparison of test set mean squared error after running DSL in both type of dataset with main dataset.

	Model Type	All	Drama	Comedy	Horror
Old Data	Pooled	5.63	5.57	**5.78**	5.55
BLS	Pooled	5.63	5.57	**5.78**	5.55
BSLS	Pooled	5.62	**5.55**	**5.78**	5.57
Old Data	separate	5.72	5.63	5.97	5.61
BLS	separate	5.72	5.63	5.97	5.61
BSLS	separate	5.70	5.61	5.95	5.59
Old Data	Data Shared	**5.54**	**5.56**	5.85	**5.07**
BLS	Data Shared	**5.54**	**5.56**	5.85	**5.07**
BSLS	Data Shared	**5.54**	5.56	5.85	**5.07**

Comparison of extracted features through venn diagram is shown in fig.(5). Following fig.(5) is showing venn diagram comparison of the features extracted after running DSL model in all three types of dataset. The sub fig.(a) is for the dataset with 27743 features, sub fig.(b) is for the dataset with 11295 features and sub fig.(c) is for the dataset with 4499 features. From the venn diagram representation we can see that feature distribution is almost same in all three types of the dataset.

We implemented our methodology (3.2) on IMDb dataset. Here we evaluate DSL model with differential regularization and the results are given in following table (3). From the table (3) we see that the average mean squared difference between true rating and predicted rating varies with the choice of the regularization parameter r_g. Selected regularization parameter r_g performs better in terms of MSE.

Following table (4) is showing the relative change in test set MSE in percentage in view of fig.(6). Here we study the effect of different subgroups on DSL model with differential regularization r_g. We implemented our methodology (3.3) on IMDb dataset. Following fig.(6) showing the venn diagram representation of the shared coefficient and separate coefficients for different regularization. Each diagram in fig.(6) is showing results obtained after imposing a different penalty over groups. Here the size of each ellipse is proportional to the number
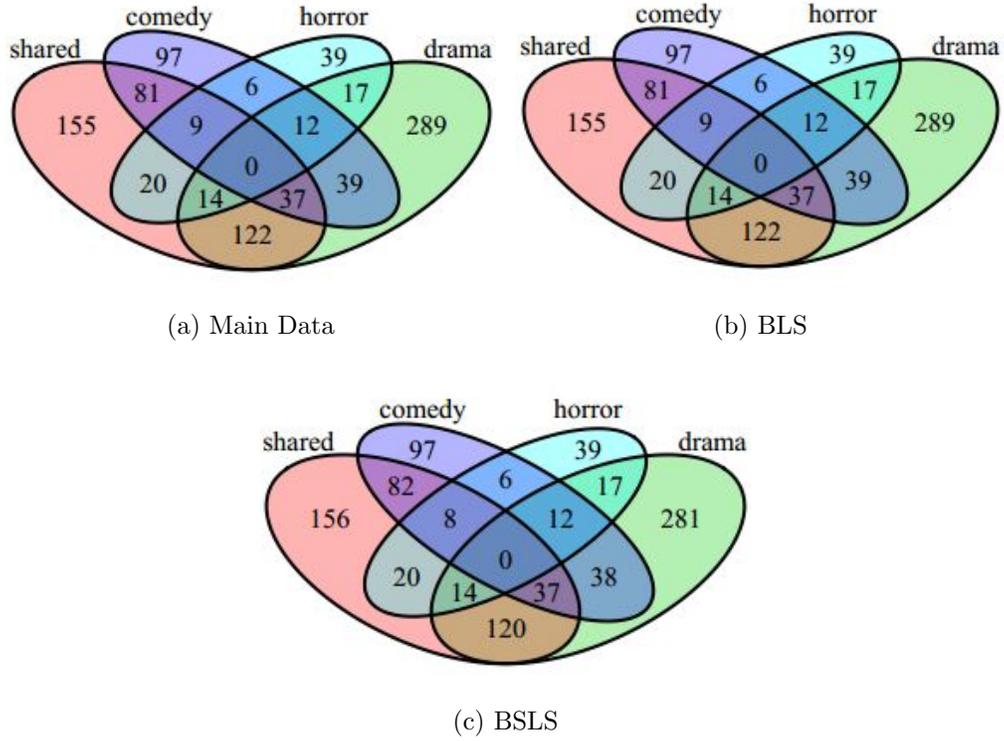
(a) Main Data

(b) BLS

(c) BSLS

FIGURE 5. The comparison of the extracted feature after running DSL model over all types of dataset.

TABLE 3. Comparison of test set MSE using different weighting scheme over groups.

Penalty type	Model	All	Drama	Comedy	Horror
$\sqrt{\frac{1}{3}}$	Data Shared	5.54	5.56	5.85	5.07
$\sqrt{\frac{n_{g_i}}{N}}$	Data Shared	5.57	5.63	5.84	5.05
$\sqrt{\frac{\log N}{\log n_{g_i}}}$	Data Shared	5.43	5.47	5.71	4.93
$\frac{N}{n_{gi}}$	Data shared	5.42	5.43	5.71	4.99
$\frac{\log N}{\log n_{gi}}$	Data shared	5.43	5.47	5.71	4.93
$\frac{\log n_{gi}}{\log N}$	Data shared	5.46	5.49	5.74	4.97
$\sqrt{\frac{\log n_{g_i}}{\log N}}$	Data shared	5.46	5.50	5.74	4.95
$\sqrt{\frac{\log n_{g_i} \times N}{\log N \times n_{g_i}}}$	Data shared	5.42	5.44	5.69	4.96

of features in that ellipse. The diagrams in fig.(6) are divided in different subgroup of features. Following table (4) is showing the relative effect of different subgroups in terms of mean squared error. Here we removed different subgroups which are in common with shared features and the relative effect in percentage is noted in table (4). From the fig.(6) we see that there are 130 features which are in common for every choice of the regularization parameter r_g and also the removal of these words increases the relative MSE around 23% which explains the importance of these features in the model. From the table (4) we see that there is no significant relative change in MSE after removing the additional features which are added to shared features which leads to noise in the sub model.

6. Discussion

In this article we present a new approach to choose reduce the size of *bag-of-words* and also an idea about choosing the regularization parameter r_g. The generalized data enriched model 2.1 can be written as

$$y_i = x_i^T \beta + x_i^T C_{g_i} \Delta_{g_i} / C_{g_i} + \epsilon_i,$$

where C_{g_i} is the variability order of the normalization column X_i. This can be written as

$$y_i = x_i^T \beta + x_i^T \gamma_{g_i} / C_{g_i} + \epsilon_i,$$

where $\gamma_{g_i} = C_{g_i} \Delta_{g_i}$ are new group effects. The penalty term simplifies to,

$$\lambda(C_0 \|\beta\|_1 + \sum_{g=1}^{G} r_g C_g \Delta_g) = \lambda(C_0 \|\beta\|_1 + \sum_{g=1}^{G} w_g \Delta_g),$$

where $w_g = r_g C_g$ are new weights over the groups.

Standardization of columns should be done on different groups and pooled groups. Under normal thresholding when $\mathcal{O}(\sqrt{n}\hat{\beta})$ are $\mathcal{O}_P(1)$ in such of case Donho and Jhonstone find for declaring global threshold (2.3) is of order $\sqrt{2log(n)}.\gamma_1.\sigma/\sqrt{n}$. Since we have sparse matrices so order may be little low of n and since order is not known we tried different choice of n. From the equation (2.3) we have

$$t_n = \sqrt{2log(n)}.\gamma_1.\sigma/\sqrt{n},$$

this can be written as

$$\sqrt{\tfrac{n}{log(n)}} = \sqrt{2}\gamma_1.\sigma/t_n = r_g.$$

Different choice of C_0 gives different choice for the regularization parameter r_g.
Let $N = \sum_{g_i=1}^{G} n_{g_i}$ then we have
1) Choosing $n = n_{g_i} = n$, $log(n) = log(N)$ and $C_0 = N/log(N)$ weights over groups becomes $\sqrt{1/3}$.
2) Choosing $n = n_{g_i}$, $log(n) = log(N)$ and $C_0 = N/log(N)$ weights over groups becomes $\sqrt{(n_{g_i}/N)}$.
3) Choosing $n = log(n_{g_i})$, $log(n) = n_{g_i}$, $C_0 = \sqrt{log(N)/N}$ weights over groups becomes $\sqrt{\frac{\log n_{g_i} \times N}{\log N \times n_{g_i}}}$.
4) Choosing $n = \log n_{g_i}$, $log(n) = log(N)$ and $C_0 = 1$ weights over groups becomes $\sqrt{\log n_{g_i} / \log N}$.
5) Choosing $n = log(N)$, $log(n) = \log n_{g_i}$ and $C_0 = 1$ weights over groups becomes $\sqrt{\log N / \log n_{g_i}}$.
By taking square of the penalties (4) and (5) we can find other two penalties. Proposed weights satisfy the condition $\sum_g r_g > 1$.

Bootstrapped LASSO and DSL can be think of a data reduction technique which reduces the size of *bag-of-features* that keeps all the important features available in the dataset. Fig. (2) is showing the word cloud for the top hundred sentiments in terms of proportion of words appeared after running Bootstrapped LASSO on each genre. The table (3) shows that the choice of the differential regularization also has improved prediction accuracy in terms of MSE. From the table (4) we see that additional features added to shared features can be assumed to be noise as there is no greater change in relative MSE. Those features which are common in different groups seems to be important sentiments as removal of these features results in the greater change in relative MSE (4) and also fig. (6) depicts that there are almost 130 features which are common in all groups as well in shared features for all types of penalty and removal of these features has greater change in relative MSE which indicates that this representation gives us smallest subset of the important sentiments.

Acknowledgement

References

[1] Bautin M., Vijayarenu L., and Skiena S., (2008). International sentiment analysis for news and blogs., *In Proceedings of the International Conference on Weblogs and Social Media (ICWSM).*

[2] Bühlmann P. and Van de Geer S., (2011). Statistics for High-Dimensional Data: Methods, Theory and Applications, Springer.

[3] Chen A., Owen A.B., and Shi M., (2013). Data enriched linear regression., *Arxiv e-prints. arxiv:1304,1837.*

[4] Dave K., Lawrence S., and Pennock D.M., (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews., *In Proceedings of the 12th international conference on World Wide Web. ACM.*, 519-528.

[5] Donoho D.L., (1995). Denoising via soft thresholding, *IEEE Transactions on Information Theory.* **41**, 613627.

[6] Efron et.al., (2011). Least Angle Regression, Annals of Statistics **32(2)**, 407499.

[7] Efron B. and Tibshirani R., (1993). An introduction to Bootstrap, *London: Chapman and Hall.*

[8] Efron B. and Tibshirani R.J., (1994). An Introduction to the Bootstrap, Chapman and Hall/CRC Monographs on Statistics and Applied Probability.

[9] Friedman J., Hastie T., and Tibshirani R., (2010). Regularization paths for generalized linear models via coordinate decent., *J.Stat.Softw.* **33**, 1.

[10] Godbole N., Srinivasaiah M., and Skiena S., (2007) Large-scale sentiment analysis for news and blogs., *In Proceedings of the International Conference on Weblogs and Social Media (ICWSM). Citeseer.*

[11] Gross S.M. and Tibshirani R., (2016). Data Shared Lasso: A novel tool to discover and uplift, Computational Statistics and Data Analysis **101**, 226-235.

[12] Hatzivassiloglou V. and McKeown K.R., (1997). Predicting the semantic orientation of adjectives., *In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics.*, 174-181.

[13] Hoerl A and Kennard R., (1988.) Ridge regression, Encyclopedia of Statistical Sciences **8**, 129136.

[14] Maas et.al., (2011). Learning Word Vectors for Sentiment Analysis, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 142–150.

[15] Mahalanobis P.C., (1946).Recent experiments in statistical sampling in the Indian Statistical Institute., Jour. Roy. Stat. Soc. **109**, 325-370.

[16] Meinshausen and Bühlmann P., (2006). High dimensional graphs and variable selection with the lasso, Annals of Statistics **34**, 1436-1462.

[17] Miller A.J., (1990). Subset Selection in Regression (3rd ed.), Chapman and Hall, London.

[18] Na J.C., Sui H., and Khoo et. al. C., (2004). Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews., *Advances in Knowledge Organization.* **9**, 49-54.

[19] Pang B., Lee L., and Vaithyanathan S., (2002). 'Thumbs up? Sentiment classification using machine learning techniques', *In Proceedings of the ACL-02 conference on Empirical methods in natural language processing. Association for Computational Linguistics.* **10**, 79-86.

[20] Quenouille M., (1949). Approximate tests of correlation in time series., Journal of the Royal Statistical Society: Series B **11**, 18-84.

[21] ———, (1956). Notes on bias in estimation., Biometrika **43**, 353-360.

[22] Tibshirani R., (1996). Regression shrinkage and selection via the lasso., *J.R.Stat.Soc.Ser.B.* **58**, 267-288.

[23] Tukey J., (1958). Bias and confidence in not quite large samples, abstract., Ann. Math. Statist. **29**, 614.

[24] Wiebe J. and Rebecca B., (1995). Probabilistic classifiers for tracking point of view., *In Proceedings of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation.*, 181-187.

[25] Zou H. and Hastie T., (2005). Regularization and variable selection via the elastic net., Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67**, 301320.

(a) $\sqrt{\frac{1}{3}}$

(b) $\sqrt{\frac{n_{g_i}}{N}}$

(c) $\sqrt{\frac{\log n_{g_i}}{\log N}}$

(d) $\sqrt{\frac{\log N}{\log n_{g_i}}}$

(e) $\frac{\log n_{g_i}}{\log N}$

(f) $\frac{\log N}{\log n_{g_i}}$

(g) $\sqrt{\frac{\log n_{g_i} \times N}{\log N \times n_{g_i}}}$
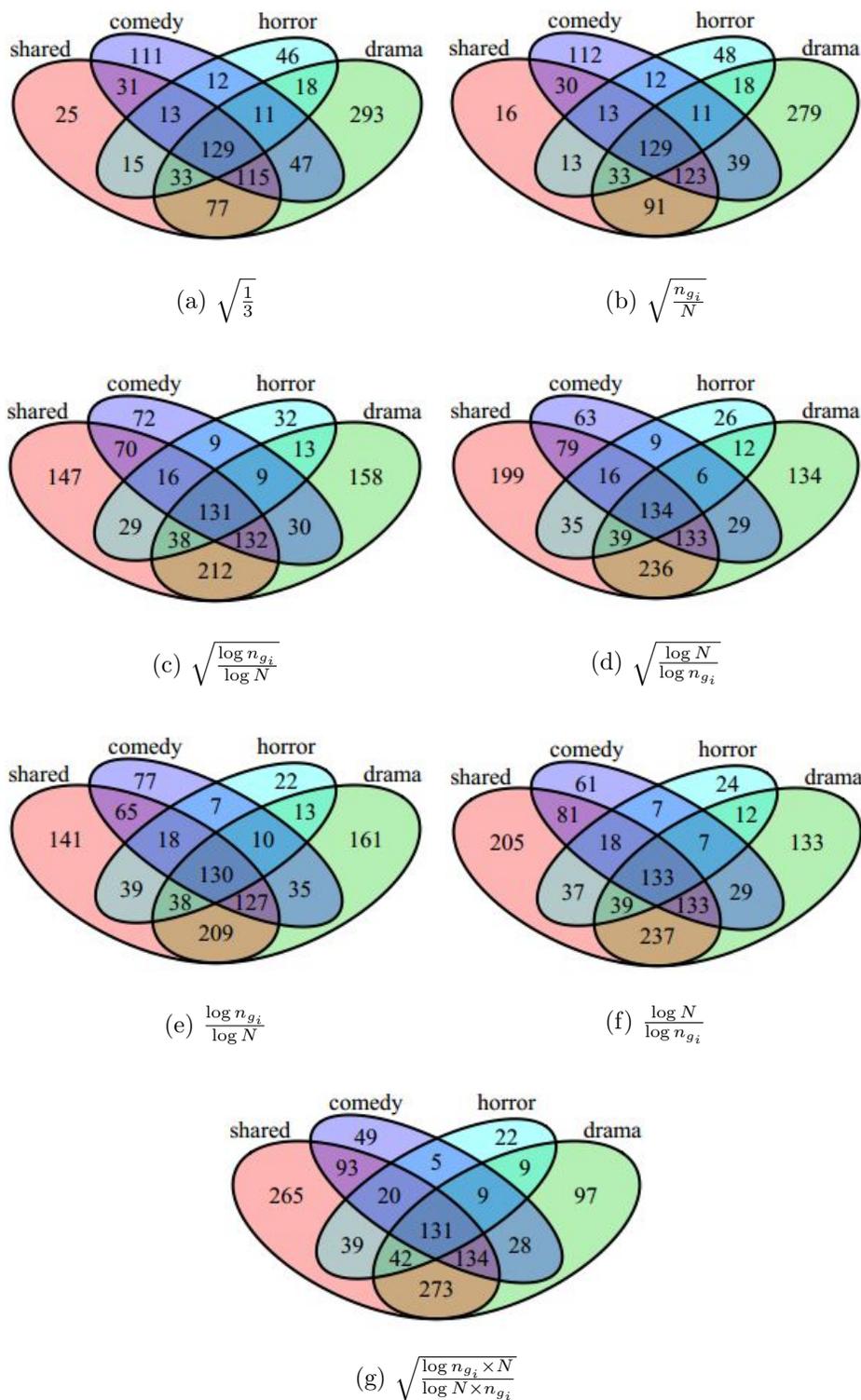
FIGURE 6. Here are the venn diagrams between the coefficient found after running LASSO on each genre and shared coefficient found after running DSL model for different values of regularization parameter r_g on an IMDb sentiment analysis dataset.

TABLE 4. Following table shows the relative change in MSE in percentage by removing different subgroup of features in view of figure (6). Following table explains the importance of the subgroups of the features based on relative change in MSE. The relative change in MSE is noted after removing the no of features (which are in right most column) from the main model.

Penalty type	Removal type	MSE % increase in All	MSE % increase in Drama	MSE % increase in Comedy	MSE % increase in Horror	No.of coef. removed
$\sqrt{\frac{1}{3}}$	no removal	0	0	0	0	0
	all intersection	25.8	23.5	31.6	23.0	129
	shared int dram	40.2	39.0	46.4	33.6	354
	shared int com	36.7	34.7	43.5	31.6	288
	shared int hor	28.4	25.9	35.2	25.1	190
	additional	0.4	0.5	0.6	0.1	25
$\sqrt{\frac{n_{g_i}}{N}}$	no removal	0	0	0	0	0
	all intersection	24.5	22.3	29.9	22.1	129
	shared int dram	40.1	39.5	45.0	33.9	376
	shared int com	36.1	34.0	42.6	31.3	295
	shared int hor	27.2	24.9	33.6	23.6	188
	additional	0.04	0.02	0.05	0.02	16
$\sqrt{\frac{\log N}{\log n_{g_i}}}$	no removal	0	0	0	0	0
	all intersection	23.9	22.3	28.7	20.7	134
	shared int dram	42.4	43.8	45.7	33.1	542
	shared int com	37.9	35.9	45.5	30.9	362
	shared int hor	26.8	25.3	32.4	21.8	224
	additional	1.1	1.3	1.2	0.7	199
$\sqrt{\frac{\log n_{g_i}}{\log N}}$	no removal	0	0	0	0	0
	all intersection	23.6	21.9	28.3	20.8	31
	shared int dram	42.0	43.2	45.9	32.1	513
	shared int com	37.0	35.4	44.1	30.5	349
	shared int hor	26.4	24.6	32.0	21.9	214
	additional	1.0	1.2	1.1	0.6	147
$\frac{\log n_{gi}}{\log N}$	no removal	0	0	0	0	0
	all intersection	23.2	21.9	27.5	20.1	130
	shared int dram	41.2	42.7	44.5	31.6	504
	shared int com	36.9	35.9	43.1	29.8	340
	shared int hor	26.2	24.8	31.3	22.1	225
	additional	1.1	1.2	1.1	0.7	141
$\frac{\log N}{\log n_{gi}}$	no removal	0	0	0	0	0
	all intersection	23.4	21.8	28.1	20.6	133
	shared int dram	42.1	43.6	45.2	32.8	542
	shared int com	37.9	35.9	45.9	30.7	365
	shared int hor	26.3	24.7	32.1	21.7	227
	additional	1.2	1.3	1.2	0.8	205
$\sqrt{\frac{\log n_{g_i} \times N}{\log N \times n_{g_i}}}$	no removal	0	0	0	0	0
	all intersection	23.2	22.1	27.7	19.2	131
	shared int dram	41.8	43.7	43.7	33.0	580
	shared int com	37.2	35.6	45.2	29.1	378
	shared int hor	26.1	25.1	31.1	20.9	232
	additional	1.8	1.9	1.9	1.1	265