Appendix A: Proof of Theorems

A.1. Proof of Theorem 1

Assume the linear regression model is denoted by

$$Y = X\beta + \varepsilon,\ W = X + \mu.$$

The detailed model specification is described in Section 3.1 in the main paper. In the model, we do not specify only one independent variable with measurement error, so the following consistent estimator can be applied to the model with more than one independent variable with measurement error.

If we directly estimate $Y = W\beta + \varepsilon$,

$$\begin{aligned}
\underset{n->\infty}{plim}\,\hat{\beta} &= \underset{n->\infty}{plim}\,(W'W)^{-1}W'Y \\
&= \underset{n->\infty}{plim}\,(W'W)^{-1}W'(W\beta - \mu\beta + \varepsilon) \\
&= \beta - \underset{n->\infty}{plim}\,(W'W)^{-1}W'(\mu\beta + \varepsilon) \\
&= \beta - \underset{n->\infty}{plim}\,(W'W)^{-1}W'(\mu\beta).
\end{aligned} \tag{1}$$

Therefore, the OLS estimator of $Y = W\beta + \varepsilon$ is inconsistent. We propose the consistent estimator:

$$\beta^* = [I - (W'W)^{-1}W'\mu]^{-1}\hat{\beta}.$$

Proof of consistency:

$$\begin{aligned}
\underset{n->\infty}{plim}\,\beta^* &= \underset{n->\infty}{plim}\,[I - (W'W)^{-1}W'\mu]^{-1}\underset{n->\infty}{plim}\,\hat{\beta} \\
&= \underset{n->\infty}{plim}\,[I - (W'W)^{-1}W'\mu]^{-1}[I - (W'W)^{-1}W'\mu]\beta = \beta.
\end{aligned} \tag{2}$$

Variance of β^*:

$$\begin{aligned}
\mathrm{Var}(\beta^*) &= \mathrm{E}(\beta^* - E(\beta^*))^2 \\
&= \mathrm{E}\{[I - (W'W)^{-1}W'\mu]^{-1}(W'W)^{-1}W'\varepsilon\}^2 \\
&= \sigma^2 \times ABW'WB^TA^T.
\end{aligned} \tag{3}$$

$$A = (I - (W'W)^{-1}W'\mu)^{-1}; B = (W'W)^{-1}.$$

For $W'\mu$, we can estimate it on a labeled set in hybrid studies. Moreover, when the sample size of labeled set is large enough, $W'\mu$ will converge to the true values.

A.2. Proof of Theorem 2

Assume the binary choice model is defined as

$$P(Y=1|X,Z)=F(X\beta+Z\gamma),\ y=Y+\mu. \tag{4}$$

The detailed specification is described in Section 3.2 in the main paper. The misclassification probabilities are denoted by

$$a_0=P(y=1|Y=0,X,Z),\ a_1=P(y=0|Y=1,X,Z).$$

The following conditional probability function of y is derived according to probability theory:

$$\begin{aligned}P(y=1|X,Z)&=P(y=1,Y=0|X,Z)+P(y=1,Y=1|X,Z)\\&=P(y=1|Y=0,X,Z)P(Y=0|X,Z)+P(y=1|Y=1,X,Z)P(Y=1|X,Z)\\&=a_0[1-F(X\beta+Z\gamma)]+(1-a_1)F(X\beta+Z\gamma)\\&=a_0+(1-a_0-a_1)F(X\beta+Z\gamma).\end{aligned} \tag{5}$$

Variance of β^* and γ^*:

The asymptotic variance-covariance matrix of the maximum likelihood estimator is defined as (Greene 2012),

$$\begin{aligned}\mathrm{Var}(\beta)&=-\mathrm{E}[\frac{\partial^2 lnL(\beta|y)}{\partial\beta\partial\beta'}]^{-1}\\&=-[\frac{1}{m}\sum_{i=1}^{m}\mathrm{E}(\frac{\partial^2\ln(\mathrm{Prob}_i)}{\partial\beta\partial\beta'})]^{-1}\times\frac{1}{m},\end{aligned} \tag{6}$$

where Prob_i is the probability function for y_i in the likelihood function $L(\beta|y)$ and m is the sample size.

We first derive the asymptotic variance-covariance matrix for the binary choice model without misclassification. To simplify the derivation, we use $F(X\beta)$ to denote $P(Y=1|X)$. Let $\mathrm{Prob}_i=y_i\ln F_i+(1-y_i)\ln[1-F_i]$, we derive

$$\begin{aligned}\mathrm{E}(\frac{\partial^2\ln(\mathrm{Prob}_i)}{\partial\beta\partial\beta'})&=\mathrm{E}[\frac{y_i f_i^{'}}{F_i}-\frac{f_i^2 y_i}{F_i^2}+\frac{f_i^{'}(y_i-1)}{1-F_i}+\frac{(y_i-1)f_i^2}{(1-F_i)^2}]x_i x_i^{'}\\&=[\frac{\mathrm{E}(y_i)f_i^{'}}{F_i}-\frac{f_i^2\mathrm{E}(y_i)}{F_i^2}+\frac{f_i^{'}(\mathrm{E}(y_i)-1)}{1-F_i}+\frac{(\mathrm{E}(y_i)-1)f_i^2}{(1-F_i)^2}]x_i x_i^{'}\\&=[\frac{F_i f_i^{'}}{F_i}-\frac{f_i^2 F_i}{F_i^2}+\frac{f_i^{'}(F_i-1)}{1-F_i}+\frac{(F_i-1)f_i^2}{(1-F_i)^2}]x_i x_i^{'}\\&=\frac{f_i^2}{(F_i-1)(F_i)}x_i x_i^{'},\end{aligned} \tag{7}$$

where f_i and $f_i^{'}$ are $\frac{\partial F_i}{\partial(x_i\beta)}$ and $\frac{\partial^2 F_i}{\partial(x_i\beta)^2}$. Then assuming n as the sample size, $\mathrm{Var}(\beta)$ is derived by

$$\begin{aligned}\mathrm{Var}(\beta)&=-[\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(\frac{\partial^2\ln(\mathrm{Prob}_i)}{\partial\beta\partial\beta'})]^{-1}\frac{1}{n},\\&=[\frac{1}{n}\sum_{i=1}^{n}\frac{f_i^2}{F_i(1-F_i)}x_i x_i^{'}]^{-1}\times\frac{1}{n}\end{aligned} \tag{8}$$

To derive the asymptotic variance-covariance matrix for the binary choice model with misclassification, we can replace F_i with P_i in Eq.(5), f_i with the first-order derivative of P_i ,

$$\text{Var}\begin{bmatrix}\gamma^*\\ \beta^*\end{bmatrix} = [\frac{1}{N}\sum_{i=1}^{N}\frac{(1-a_{i0}-a_{i1})^2 f_i^2}{P_i(1-P_i)}\begin{bmatrix}z_i\\ x_i\end{bmatrix}\begin{bmatrix}z_i & x_i\end{bmatrix}]^{-1}\times\frac{1}{N}, \tag{9}$$

where N is the sample size, P_i is defined by Eq.(5), and f_i is the derivative of F_i in Eq.(4).

A.3. Proof of Theorem 3

Assume the generalized linear model is defined as

$$P(Y|X,Z) = G(X\beta + Z\gamma),\ W = X + \mu. \tag{10}$$

The detailed specification is described in Section 3.3 in the main paper. The conditional probability function is derived by probability theory. Assume that W provides no additional information about Y conditional on X:

$$\begin{aligned}
&P(Y=y|W=w,Z=z)\\
&=\sum_X P(Y=y,X=x|W=w,Z=z)\\
&=\sum_X P(Y=y|X=x,W=w,Z=z)P(X=x|W=w,Z=z)\\
&=\sum_X P(Y=y|X=x,Z=z)P(X=x|W=w,Z=z)\\
&=P(Y=y|X=0,Z=z)P(X=0|W=w,Z=z)+P(Y=y|X=1,Z=z)P(X=1|W=w,Z=z)\\
&=G(0\times\beta+Z\gamma)[1-F(W\alpha+Z\theta)]+G(1\times\beta+Z\gamma)F(W\alpha+Z\theta).
\end{aligned} \tag{11}$$

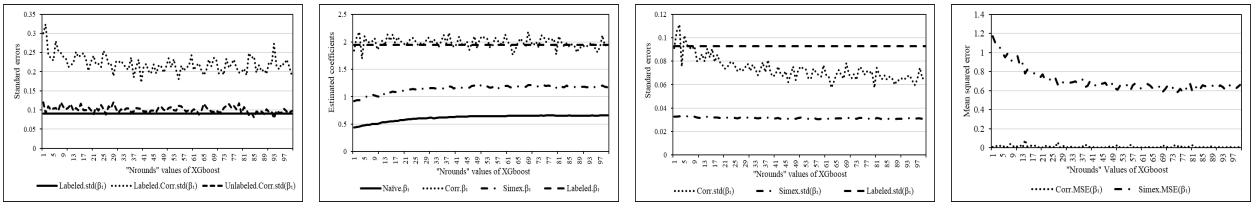
Variance of β^* and γ^*:

Since GLM has many cases, we take the binary choice model as an example to derive the asymptotic variance-covariance matrix, following the derivation process in Eq.(7) and Eq.(8),

$$\text{Var}\begin{bmatrix}\gamma^*\\ \beta^*\end{bmatrix} = [\frac{1}{N}\sum_{i=1}^{N}\frac{(a_{i1}g_{i1}\begin{bmatrix}z_i\\ 1\end{bmatrix}+a_{i0}g_{i0}\begin{bmatrix}z_i\\ 0\end{bmatrix})(a_{i1}g_{i1}\begin{bmatrix}z_i & 1\end{bmatrix}+a_{i0}g_{i0}\begin{bmatrix}z_i & 0\end{bmatrix})}{P_i(1-P_i)}]^{-1}\times\frac{1}{N}, \tag{12}$$

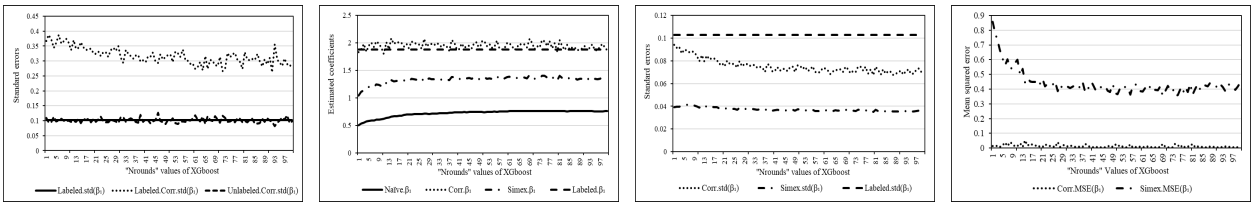
where G_i is defined by Eq.(10), P_i is defined by Eq.(11), g_i is the derivative of G_i, $g_{i1} = g(1\times\beta + z_i\gamma)$, $g_{i0} = g(0\times\beta + z_i\gamma)$, $a_{i1} = F(w_i\alpha + z_i\theta)$ and $a_{i0} = 1 - F(w_i\alpha + z_i\theta)$.

Figure 1 Simulation Results for Case 2 (Probit Model)



(a) Proportion Results (b) Coefficient Results (c) Variance Results (d) MSE Results

Figure 2 Simulation Results for Case 3 (Probit Model)



(a) Proportion Results (b) Coefficient Results (c) Variance Results (d) MSE Results

Appendix B: Additional Results for Amazon Review Data Set

In this subsection, the results for the probit models in Case 2 and Case 3 are reported in Figure 1 and Figure 2, respectively. Due to the page limit, we only report the results for the focal variable in this and next subsections. The results are qualitatively similar to our main results.

We also conducted three additional experiments on the review dataset by varying the classifier parameter, classifier type, and base rate of the dataset. Specifically, we conducted the experiments by tuning the parameter of XGBoost, "max_depth", from four to eight with step value equaling 1. "Max_depth" indicates the maximum depth of a tree. The results are shown in Table 1. We also conducted the experiments by tuning the parameter of Radial Basis Function (RBF) kernel SVM, "C", from 0.5 to 1.5 with step value equaling 0.25. The C parameter trades off the correct classification of training examples against maximization of the decision function's margin. The results are shown in Table 2. Finally, we conducted the experiments after we change the base rate of the review dataset to 10% of the dataset as negative reviews and 90% of the dataset as positive reviews. We also tune the parameter of XGBoost, "max_depth", from four to eight with step value equaling 1. The results are shown in Table 3. The theoretical coefficients of β_1 and β_2 are still 2 and 1. The results show that our method still corrects the coefficients around the theoretical values. MC-SIMEX still only partially corrects the inconsistency.

Table 1 Coefficient Results for "Max_depth" of XGBoost

	XGB_depth	Naive.β_1	Naive.β_2	Corr.β_1	Corr.β_2	Simex.β_1	Simex.β_2
Case 1	4	0.699	1.333	2.042	1.009	1.37	1.248
	5	0.725	1.328	1.952	1.026	1.369	1.24
	6	0.733	1.326	1.897	1.028	1.372	1.236
	7	0.737	1.325	1.925	1.022	1.372	1.233
	8	0.753	1.323	2.07	0.993	1.434	1.227
Case 2	4	0.665	0.292	2.179	1.029	1.305	0.583
	5	0.664	0.324	1.957	1.001	1.263	0.618
	6	0.644	0.337	1.808	0.999	1.199	0.643
	7	0.668	0.337	1.96	1.026	1.253	0.623
	8	0.665	0.348	1.933	1.051	1.276	0.674
Case 3	4	0.693	1.239	2.102	1.029	1.368	1.178
	5	0.719	1.236	2.043	1.041	1.37	1.174
	6	0.74	1.235	2.021	1.047	1.393	1.173
	7	0.745	1.233	2.057	1.044	1.403	1.168
	8	0.725	1.233	2.112	1.036	1.401	1.166

Appendix C: Simulation Results for Toxic Comment Data Set

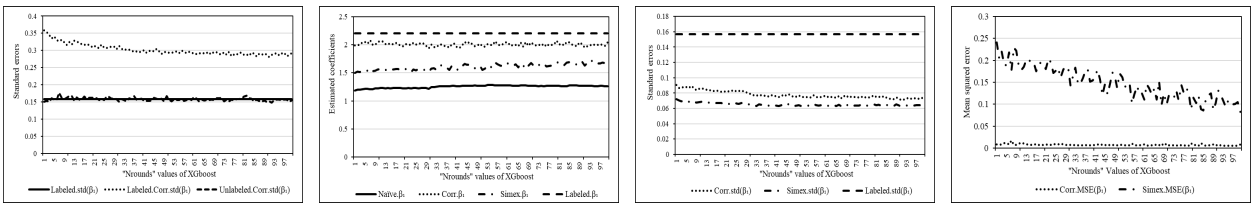
In this subsection, we report the simulation results for toxic comment data. Specifically, results for the linear regression in Case 1 are reported in Figure 3; Results for the logit and probit models in Case 2 are reported in Figure 4 and Figure 5, respectively; Results for the logit and probit model in Case 3 are reported in Figure 6 and Figure 7, respectively. The results provide a robust test for our solutions, which are qualitatively the same as our main results.

Table 4 reports the performance metric comparison results for toxic comment data. In summary, our performance metric can achieve the best precision and small bias among all the performance metrics. Among the traditional metrics, Accuracy, F1-measure, Kappa, and MCC can achieve similar performance to our

Table 2 Coefficient Results for "C" of SVM

	SVM_C	Naive.β_1	Naive.β_2	Corr.β_1	Corr.β_2	Simex.β_1	Simex.β_2
Case 1	0.5	0.768	1.334	2.131	0.962	1.468	1.231
	0.75	0.785	1.331	1.975	1.022	1.436	1.23
	1	0.793	1.329	2.017	1.004	1.462	1.225
	1.25	0.78	1.33	2.014	0.988	1.439	1.226
	1.5	0.784	1.33	1.959	1.012	1.409	1.229
Case 2	0.5	0.692	0.335	2.128	1.02	1.322	0.63
	0.75	0.696	0.353	1.908	0.966	1.284	0.652
	1	0.71	0.364	1.992	0.98	1.32	0.671
	1.25	0.721	0.364	2.148	1.05	1.357	0.685
	1.5	0.721	0.361	1.989	0.977	1.332	0.662
Case 3	0.5	0.759	1.244	2.149	1.046	1.426	1.18
	0.75	0.762	1.242	1.997	1.063	1.402	1.18
	1	0.781	1.241	2.047	1.053	1.437	1.177
	1.25	0.776	1.241	2.119	1.058	1.425	1.175
	1.5	0.78	1.241	2.023	1.063	1.405	1.177

Figure 3 Simulation Results for Case 1



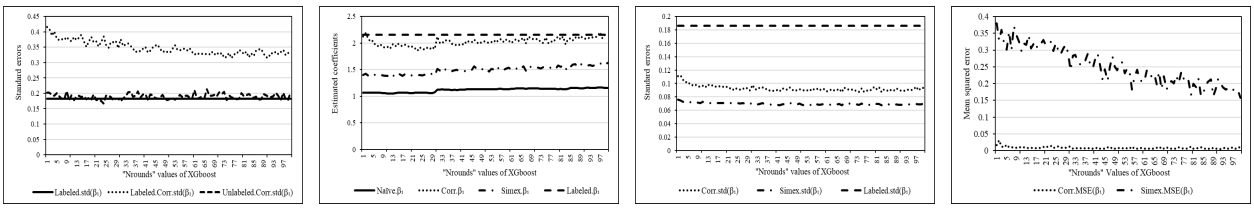
(a) Proportion Results (b) Coefficient Results (c) Variance Results (d) MSE Results

metric. Balanced accuracy, sensitivity, specificity, NPV, PPV, informedness, markedness, AUC, and Brier score perform worse compared with our metric.

Table 3 Coefficient Results when Base Rate of Negative Reviews is 10%

	XGB_depth	Naive.β_1	Naive.β_2	Corr.β_1	Corr.β_2	Simex.β_1	Simex.β_2
Case 1	4	0.196	1.181	2.298	1.008	0.425	1.172
	5	0.193	1.181	2.127	1.036	0.433	1.17
	6	0.193	1.181	1.746	1.036	0.405	1.171
	7	0.222	1.179	1.88	1.037	0.472	1.167
	8	0.237	1.179	2.187	0.988	0.526	1.166
Case 2	4	0.22	0.084	2.005	1.037	0.454	0.168
	5	0.227	0.106	2.125	1.029	0.501	0.226
	6	0.252	0.099	2.167	1.079	0.548	0.221
	7	0.273	0.1	1.934	1.014	0.597	0.206
	8	0.269	0.109	2.183	1.224	0.575	0.235
Case 3	4	0.339	1.939	2.049	1.05	0.742	1.861
	5	0.348	1.934	2.113	1.021	0.767	1.849
	6	0.342	1.935	1.856	1.125	0.739	1.852
	7	0.346	1.929	1.889	1.113	0.714	1.84
	8	0.361	1.929	1.986	1.071	0.781	1.837

Figure 4 Simulation Results for Case 2 (Logit Model)

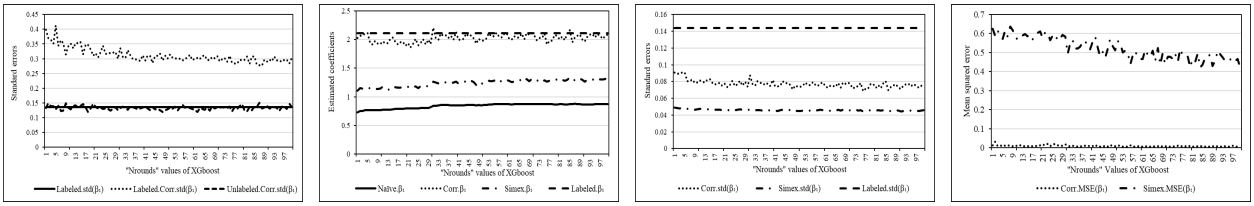


(a) Proportion Results (b) Coefficient Results (c) Variance Results (d) MSE Results

Appendix D: Additional Results for Applications to Real-World Data Set

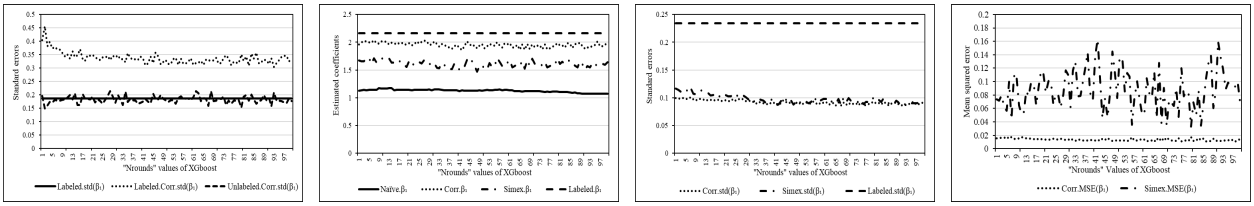
In this subsection, the coefficient results for the probit model are reported in Figure 8. Moreover, the variance results of sentiment for the three models are reported in Figure 9. The results are qualitatively the same as those in the main experiments.

Figure 5 Simulation Results for Case 2 (Probit Model)



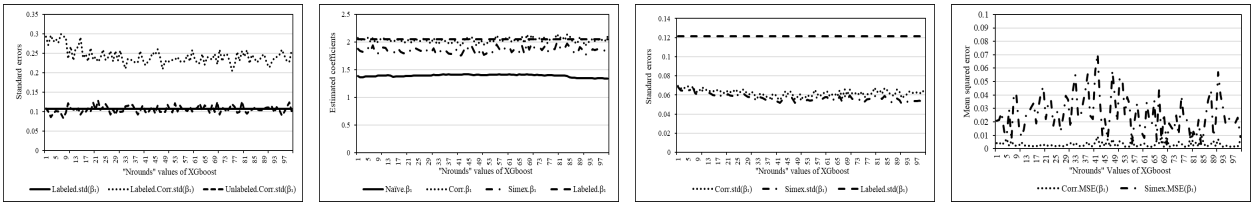
(a) Proportion Results (b) Coefficient Results (c) Variance Results (d) MSE Results

Figure 6 Simulation Results for Case 3 (Logit Model)



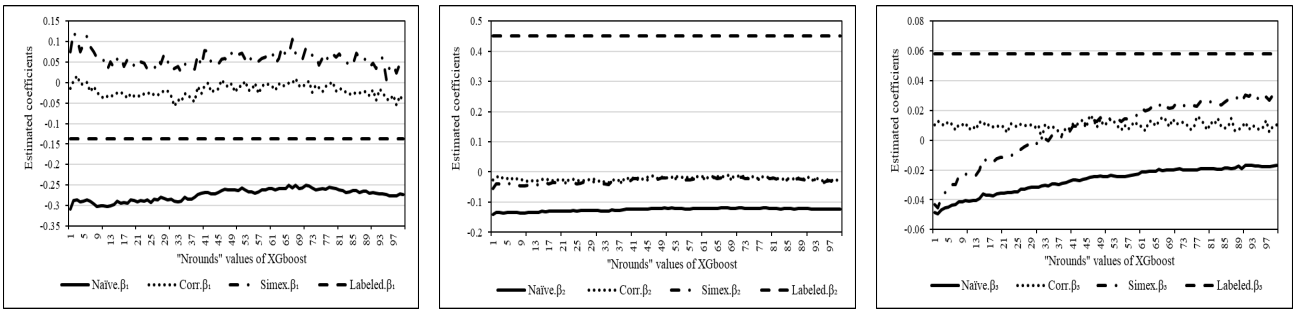
(a) Proportion Results (b) Coefficient Results (c) Variance Results (d) MSE Results

Figure 7 Simulation Results for Case 3 (Probit Model)



(a) Proportion Results (b) Coefficient Results (c) Variance Results (d) MSE Results

Figure 8 Results for Probit Model

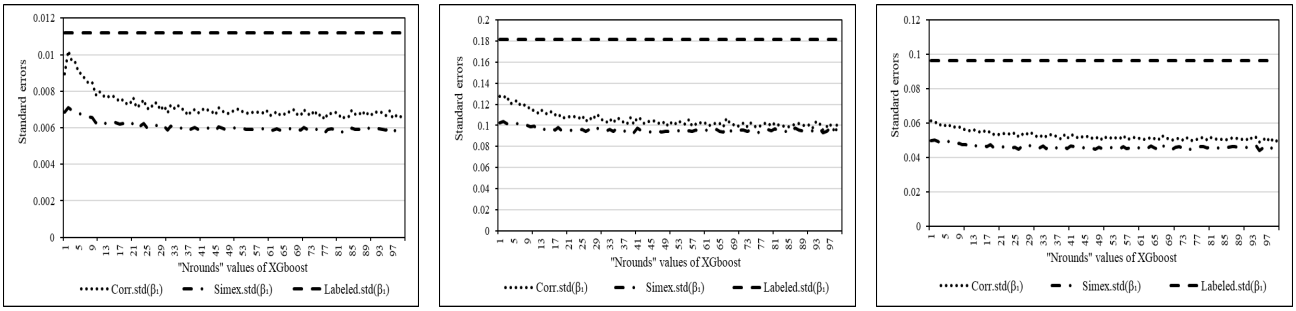


(a) Results for Sentiment (b) Results for No of Votes (c) Results for Word Count

Table 4 Performance Metric Comparison Results for Toxic Comment Dataset

	Case 1			Case 2			Case 3		
	$Bias^2$	Variance	Nrounds	$Bias^2$	Variance	Nrounds	$Bias^2$	Variance	Nrounds
Min.std.error	0.0003	0.0061	282	0.0000	0.0084	259	0.0002	0.0092	179
Accuracy	0.0003	0.0062	276	0.0000	0.0084	276	0.0000	0.0097	276
F1	0.0003	0.0062	276	0.0000	0.0084	276	0.0000	0.0097	276
Kappa	0.0003	0.0062	276	0.0000	0.0084	276	0.0000	0.0097	276
MCC	0.0003	0.0062	276	0.0000	0.0084	276	0.0000	0.0097	276
Sensitivity	0.0135	0.0068	267	0.0029	0.009	267	0.0023	0.0103	267
Specificity	0.0049	0.0074	48	0.0033	0.0096	48	0.0033	0.0098	48
Bacc	0.0135	0.0068	267	0.0029	0.009	267	0.0023	0.0103	267
Informedness	0.0135	0.0068	267	0.0029	0.009	267	0.0023	0.0103	267
PPV	0.0049	0.0074	48	0.0033	0.0096	48	0.0033	0.0098	48
NPV	0.0135	0.0068	267	0.0029	0.009	267	0.0023	0.0103	267
Markedness	0.0049	0.0074	48	0.0033	0.0096	48	0.0033	0.0098	48
AUC	0.0061	0.0067	252	0.0048	0.0092	252	0.0019	0.0102	252
Brier	0.0034	0.0064	300	0.0042	0.0091	300	0.0013	0.0101	300

Figure 9 Variance Results for Sentiment



(a) Results for Linear Model (b) Results for Logit Model (c) Results for Probit Model