Appendices of Model Evaluation with Precision, Recall, and f_1 Measure Based on Block-regularized $m\times 2$ Cross Validation for Text Corpus

Outline. The monte-carlo calculation of probability $P(H_0)$ is presented in Appendix A. The average estimators of p, r, and f_1 measure based on $m \times 2$ BCV are shown in Appendix B. Several lemmas related to Theorem 1 are recalled in Appendix C. The proof of Theorem 1 is provided in Appendix D, while Appendix E is dedicated to the proof of Theorem 2. Finally, Appendix F presents the detailed information of tasks and datasets.

A Monte-Carlo Calculation of Probability $P(H_0)$

In the Bayes test (i.e., Eq. (4)), the probability $P(H_0)$ of model $\mathcal{A}$ outperforming model $\mathcal{B}$ be calculated using a Monte-Carlo simulation, as shown in the following formula.

$$\int_0^1 \int_0^1 \mathbb{I}(\nu_n^{\mathcal{A}} > \nu_n^{\mathcal{B}}) \cdot f_{\mathcal{A}}(\nu_n^{\mathcal{A}}) f_{\mathcal{B}}(\nu_n^{\mathcal{B}}) d\nu_n^{\mathcal{A}} d\nu_n^{\mathcal{B}} \approx \frac{1}{L} \sum_{i=1}^{L} \mathbb{I}(s_{i,\mathcal{A}} > s_{i,\mathcal{B}}), \quad (10)$$

where $\mathbb{I}(\cdot)$ is the indicator function with a value of one if and only if the condition is true or zero otherwise. $\{s_{i,\mathcal{A}}\}_{i=1}^{L}$ and $\{s_{i,\mathcal{B}}\}_{i=1}^{L}$ denote the observations sampled from the density functions $f_{\mathcal{A}}(\cdot)$ and $f_{\mathcal{B}}(\cdot)$ of metrics $\nu_n^{\mathcal{A}}$ and $\nu_n^{\mathcal{B}}$, L is the number of samples, and $L = 1,000,000$ is used because a large L indicates a highly accurate Monte-Carlo simulation.

B Average Estimators of p, r, and f_1 Measure Based on $m \times 2$ BCV

The average estimatiors of p, r, and f_1 measure are proposed based on the average confusion matrix from $m \times 2$ BCV, which is introduced by [8]. Before presenting a detailed introduction of these average estimators, $m \times 2$ BCV is first recalled.

The $m\times 2$ BCV is the abbreviation of block-regularized $m\times 2$ cross validation and also a specific version of $m \times 2$ CV that satisfies intra-group and inter-group regularization constraints [9]. The $m\times 2$ CV refers to m replicated data splittings of two-fold CV. The detailed review of intra-group and inter-group constraints is presented as follows.

(1) **Intra-group constraint**. The frequency distribution over the linguistic units of the training set is identical to that of the validation set in each two-fold CV of $m \times 2$ BCV.

(2) **Inter-group constraint**. The number of overlapping samples between the training sets (validation sets) from any two data splittings in an $m \times 2$ BCV is approximately equal.

On the basis of $m \times 2$ BCV, the average estimatiors of p, r, and f_1 measure are recalled as follows. Specifically, let D_n denote a text corpus, where n is the number of labeled units in D_n. We denote $\mathcal{H}(n)$ as the theoretical confusion matrix derived from the entire dataset D_n. Let $\{(\mathcal{H}_1^{(i)}, \mathcal{H}_2^{(i)})\}_{i=1}^m$ be a collection of confusion matrices from the i-th two-fold CV of $m \times 2$ BCV, where $\mathcal{H}_k^{(i)} \triangleq (TP_k^{(i)}, FP_k^{(i)}, FN_k^{(i)}, TN_k^{(i)})$. Then, the average estimators of p, r, f_1 measure from the average confusion matrix $\mathcal{H}_a(D_n) = (TP_a, FP_a, FN_a, TN_a)$ are defined:

$$P_a = \frac{TP_a}{TP_a + FP_a}, \qquad R_a = \frac{TP_a}{TP_a + FN_a}, \qquad F_{1,a} = \frac{2 \cdot P_a \cdot R_a}{P_a + R_a}, \tag{11}$$

where

$$(TP_a, FP_a, FN_a, TN_a) = \frac{\omega}{2m} \sum_{i=1}^{m} \sum_{k=1}^{2} (TP_k^{(i)}, FP_k^{(i)}, FN_k^{(i)}, TN_k^{(i)}).$$

In the above equation, $\omega = 1$ represents the macro-average, $\omega = 2$ represents the micro-average, and $\omega = \frac{2m}{1+\rho_1+2(m-1)\rho_2}$ ($\omega = 2.2128$ when $m = 3$, and the integration operation of ρ_1 and ρ_2 is performed on the intervals of $0 \leq \rho_1 \leq 0.5, 0.25 \leq \rho_2 \leq 0.5$) represents the weighted average based on the effective confusion matrix proposed by [8], where ρ_1 and ρ_2 represent intra-group and inter-group correlations, and $0 \leq \rho_1, \rho_2 < 1$.

Unfortunately, the estimation of ρ_2 is still a challenging problem. Therefore, it is necessary to avoid estimating ρ_2. The proposed voting aggregation estimators of p, r, and f_1 measure provide a suitable solution.

C Several Lemmas Related to Theorem 1

To study the theoretical properties of the voting aggregation estimators P_m, R_m, $F_{1,m}$, we first introduce Lemma 1 to provide the distribution of the theoretical confusion matrix $\mathcal{H}(n)$ defined on the dataset D_n, which is essential for deriving the distribution of the voting confusion matrix $\mathcal{H}_m(D_n)$. Then, we derive the distributions of the components both the voting confusion matrix $\mathcal{H}_m(D_n)$ and its limiting form $\mathcal{H}_\infty(D_n)$ as $m \to \infty$.

Lemma 1. *[5] The confusion matrix $\mathcal{H}(n) = (TP, FP, FN, TN)$ follows a multinomial distribution, denoted as $\mathcal{H}(n) \sim \mathcal{M}(n, \boldsymbol{\pi})$. Here, the parameter vector $\boldsymbol{\pi} \equiv (\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00})$ satisfies $\pi_{11} + \pi_{10} + \pi_{01} + \pi_{00} = 1$, with the components defined as : $\pi_{11} \equiv P(z = 1, l = 1)$, $\pi_{10} \equiv P(z = 1, l = 0)$, $\pi_{01} \equiv P(z = 0, l = 1)$, and $\pi_{00} \equiv P(z = 0, l = 0)$. Moreover, let $\pi_1 \equiv \pi_{11} + \pi_{01}$ and $\pi_0 \equiv \pi_{10} + \pi_{00}$ represent the prior probabilities of class labels with $l = 1$ and $l = 0$ in dataset D_n respectively.*

Proof. The Lemma 1 is similar to Assumption 1 of [5], and the proof is omitted here.

From Lemma 1, the counts of class labels with $l = 1$ and $l = 0$ in dataset D_n are denoted as $TP + FN = n\pi_1$ and $TN + FP = n\pi_0$ respectively, both of which are constants. Under this condition, the components TP, FP, FN, and TN follow the binomial distributions.

Property 1.

$$\begin{aligned}
&TP|TP + FN \sim Binom(n\pi_1, p_1 = \tfrac{\pi_{11}}{\pi_1}),\\
&FN|TP + FN \sim Binom(n\pi_1, 1 - p_1 = \tfrac{\pi_{01}}{\pi_1}),\\
&TN|TN + FP \sim Binom(n\pi_0, p_0 = \tfrac{\pi_{00}}{\pi_0}),\\
&FP|TN + FP \sim Binom(n\pi_0, 1 - p_0 = \tfrac{\pi_{10}}{\pi_0})\\
&E(TP) = n\pi_1 p_1, \;\; E(FN) = n\pi_1(1 - p_1),\\
&E(TN) = n\pi_0 p_0, \;\; E(FP) = n\pi_0(1 - p_0),
\end{aligned}$$

where $Binom(\cdot, \cdot)$ denotes the binomial distribution.

Proof. From Lemma 1, the counts of class labels with $l = 1$ and $l = 0$ are denoted as $TP + FN \equiv n\pi_1$ and $TN + FP \equiv n\pi_0$ respectively, both of which are constants given the dataset D_n.

Given these counts, $TP|TP + FN$ follows a binomial distribution with parameters $n\pi_1$ and $p_1 = \frac{\pi_{11}}{\pi_1}$. Similarly, $FN|TP + FN$, $TN|TN + FP$ and $FP|TN + FP$ also follow the binomial distributions, that is,

$$\begin{aligned}
&TP|TP + FN \sim Binom(n\pi_1, p_1 = \tfrac{\pi_{11}}{\pi_1}),\\
&FN|TP + FN \sim Binom(n\pi_1, 1 - p_1 = \tfrac{\pi_{01}}{\pi_1}),\\
&TN|TN + FP \sim Binom(n\pi_0, p_0 = \tfrac{\pi_{00}}{\pi_0}),\\
&FP|TN + FP \sim Binom(n\pi_0, 1 - p_0 = \tfrac{\pi_{10}}{\pi_0}).
\end{aligned}$$

Then, since the counts $TP + FN \equiv n\pi_1$ and $TN + FP \equiv n\pi_0$ are constants, the expectations of TP, FN, TN, and FP can be expressed as follows:

$$\begin{aligned}
&E(TP) = n\pi_1 p_1, \;\; E(FN) = n\pi_1(1 - p_1),\\
&E(TN) = n\pi_0 p_0, \;\; E(FP) = n\pi_0(1 - p_0).
\end{aligned}$$

For a binary classification model, in $m \times 2$ BCV, the collection of m groups of predicted outcomes of size n be represented as a binary matrix $\mathcal{Z} = (z_{i,j})_{m\times n}$ of size $m \times n$. Here, $z_{i,j}$ represents the predicted outcome of the j-th sample in D_n within the i-th two-fold cross validation, and it is considered as a random variable with the following Bernoulli distribution, i.e.,

$$\begin{aligned}
&z_{i,j} \sim Bernoulli(1, p_1), && \forall j \in \{j : l_j = 1\},\\
&z_{i,j} \sim Bernoulli(1, 1 - p_0), && \forall j \in \{j : l_j = 0\}.
\end{aligned} \tag{12}$$

Then, the predicted outcome based majority voting $v_{j,m}$ of j-th sample can be derived by the average of m predicted outcomes $\{z_{1,j},\ z_{2,j},\ \cdots,\ z_{m,j}\}$ of

j-th sample, that is,

$$v_{j,m} = \begin{cases} 1, & \frac{1}{m}\sum_{i=1}^{m} z_{i,j} \geq 0.5 \\ 0, & \frac{1}{m}\sum_{i=1}^{m} z_{i,j} < 0.5 \end{cases}. \tag{13}$$

For the j-th sample, the predicted outcome $v_{j,m}$ can be regarded as a random variable following a Bernoulli distribution, where the parameter is determined by the distribution of the average predicted outcome $\frac{1}{m}\sum_{i=1}^{m} z_{i,j}$. In fact, [14] has established the following beta distributions for the average predicted outcome $\frac{1}{m}\sum_{i=1}^{m} z_{i,j}$ in two cases of j-th sample, where $j \in \{j : l_j = 1\}$ and $j \in \{j : l_j = 0\}$.

Lemma 2. – *If* $j \in \{j : l_j = 1\}$, *then*

$$\frac{1}{m}\sum_{i=1}^{m} z_{i,j} \sim Beta((m_e - 1)p_1, (m_e - 1)(1 - p_1)); \tag{14}$$

– *If* $j \in \{j : l_j = 0\}$, *then*

$$\frac{1}{m}\sum_{i=1}^{m} z_{i,j} \sim Beta((m_e - 1)(1 - p_0), (m_e - 1)p_0), \tag{15}$$

where $m_e = \frac{m}{1+(m-1)\rho_2}$ *and* $\rho_2 = Corr(z_{i,j}, z_{i',j})$ *denotes the correlation coefficient between two different predicted outcomes* $z_{i,j}$ *and* $z_{i',j}$ *in* $m \times 2$ *BCV for* $i \neq i'$ *and* $\forall j = 1, 2, \cdots, n$. *For the symbol simplicity, let's assume that* ρ_2 *is equal on all* j.

Proof. Recall that the predicted outcome $z_{i,j}$ follows the Bernoulli distribution with the parameter p_1 in Eq. (12) for any $j \in \{j : l_j = 1\}$. Then, the expectation and variance of $z_{i,j}$ can be expressed as $E(z_{i,j}) = p_1$ and $Var(z_{i,j}) = p_1(1 - p_1)$ respectively. On this basis, we can derive the expectation and variance of the average predicted outcome $\frac{1}{m}\sum_{i=1}^{m} z_{i,j}$ as follows.

$$\begin{aligned} E\left(\frac{1}{m}\sum_{i=1}^{m} z_{i,j}\right) &= \frac{1}{m}\sum_{i=1}^{m} E(z_{i,j}) = \frac{1}{m}\sum_{i=1}^{m} p_1 = p_1, \\ Var\left(\frac{1}{m}\sum_{i=1}^{m} z_{i,j}\right) &= \frac{1}{m^2}\Big[\sum_{i=1}^{m} Var(z_{i,j}) + \sum_{i\neq i'} Cov(z_{i,j}, z_{i',j})\Big] \\ &= \frac{1}{m^2}\big[m + m(m-1)\rho_2\big]p_1(1 - p_1) \\ &= \frac{p_1(1 - p_1)}{m_e}, \end{aligned} \tag{16}$$

where $m_e = \frac{m}{1+(m-1)\rho_2}$.

Assume $\frac{1}{m}\sum_{i=1}^{m} z_{i,j}$ follows a Beta distribution $Beta(a,b)$. The parameters a and b can be determined by the expectation and variance of $\frac{1}{m}\sum_{i=1}^{m} z_{i,j}$, which are given by the following equations:

$$\begin{aligned} E\left(\frac{1}{m}\sum_{i=1}^{m} z_{i,j}\right) &= \frac{a}{a+b} = p_1, \\ Var\left(\frac{1}{m}\sum_{i=1}^{m} z_{i,j}\right) &= \frac{ab}{(a+b)^2(a+b+1)} = \frac{p_1(1-p_1)}{m_e}. \end{aligned} \tag{17}$$

Thus, the average predicted outcome $\frac{1}{m}\sum_{i=1}^{m} z_{i,j}$ follows the Beta distribution with parameters $a = (m_e - 1)p_1$ and $b = (m_e - 1)(1 - p_1)$. Similarly, for any $j \in \{j : l_j = 0\}$, we can derive the expectation and variance of the average predicted outcome $\frac{1}{m}\sum_{i=1}^{m} z_{i,j}$ as follows.

Building on Lemma 2, $v_{j,m}$ can be regarded as a random variable following a Bernoulli distribution where the parameter is deduced from the next lemma.

Lemma 3.
- *For $\forall j \in \{j : l_j = 1\}$, we have $v_{j,m} \sim Bernoulli(1, p_{1,m})$ where the parameter $p_{1,m}$ is defined as $p_{1,m} \equiv \int_{0.5}^{1} Beta((m_e - 1)p_1, (m_e - 1)(1 - p_1))dx$.*
- *For $\forall j \in \{j : l_j = 0\}$, we have $v_{j,m} \sim Bernoulli(1, 1 - p_{0,m})$ where the parameter $p_{0,m}$ is defined as $p_{0,m} \equiv \int_{0}^{0.5} Beta((m_e - 1)(1 - p_0), (m_e - 1)p_0)dx$.*
- *If $p_1 > 0.5$, then $p_{1,m} > p_1$ and $p_{1,m}$ is increasing with regard to m; and if $p_0 > 0.5$, then $p_{0,m} > p_0$ and $p_{0,m}$ is increasing with regard to m.*

Proof. Recall that the predicted outcome $v_{j,m}$ is equal to 0 or 1 and is defined as follows:

$$v_{j,m} = \begin{cases} 1, & \frac{1}{m}\sum_{i=1}^{m} z_{i,j} \geq 0.5 \\ 0, & \frac{1}{m}\sum_{i=1}^{m} z_{i,j} < 0.5 \end{cases}.$$

Then, the predicted outcome $v_{j,m}$ can be regarded as a random variable following the Bernoulli distribution with the parameter $p_{1,m}$ for $j \in \{j : l_j = 1\}$, and with the parameter $1 - p_{0,m}$ for $j \in \{j : l_j = 0\}$, respectively. The parameters $p_{1,m}$ and $p_{0,m}$ are defined as follows:

$$\begin{aligned} p_{1,m} &= P\Big(v_{j,m} = 1 \Big| j \in \{j : l_j = 1\}\Big) = P\left(\frac{1}{m}\sum_{i=1}^{m} z_{i,j} \geq 0.5 \Big| j \in \{j : l_j = 1\}\right) \\ &= \int_{0.5}^{1} Beta\big((m_e - 1)p_1, (m_e - 1)(1 - p_1)\big)dx. \\ p_{0,m} &= P\Big(v_{j,m} = 0 \Big| j \in \{j : l_j = 0\}\Big) = P\left(\frac{1}{m}\sum_{i=1}^{m} z_{i,j} < 0.5 \Big| j \in \{j : l_j = 0\}\right) \\ &= \int_{0}^{0.5} Beta\big((m_e - 1)(1 - p_0), (m_e - 1)p_0\big)dx. \end{aligned}$$

For the monotonic increasing property of $p_{1,m}$ and $p_{0,m}$ with respect to m, the results can be obtained from Table 2.

Table 2: The $p_{1,m}$ and $p_{0,m}$ for the different m, ρ_2, and p_1, p_0.

	p_1, p_0 =0.9					p_1, p_0 =0.55				
m	$\rho_2 = 0.05$	0.2	0.5	0.7	0.9	$\rho_2 = 0.05$	0.2	0.5	0.7	0.9
3	0.9418	0.9268	0.9086	0.9028	0.9003	0.5662	0.5598	0.5530	0.5510	0.5501
5	0.9679	0.9430	0.9132	0.9041	0.9004	0.5806	0.5667	0.5547	0.5514	0.5501
7	0.9807	0.9517	0.9156	0.9048	0.9005	0.5912	0.5710	0.5555	0.5517	0.5502
9	0.9875	0.9570	0.9170	0.9052	0.9005	0.5995	0.5738	0.5561	0.5518	0.5502
11	0.9914	0.9605	0.9180	0.9055	0.9005	0.6061	0.5758	0.5564	0.5519	0.5502
13	0.9938	0.9630	0.9186	0.9057	0.9005	0.6116	0.5773	0.5567	0.5520	0.5502
15	0.9953	0.9649	0.9192	0.9058	0.9006	0.6162	0.5785	0.5569	0.5520	0.5502
∞	0.9999	0.9773	0.9227	0.9068	0.9006	0.6722	0.5880	0.5582	0.5524	0.5502

Remark 2. The conditions $p_1 > 0.5$ and $p_0 > 0.5$ in Lemma 3 indicate that the model trained in each two-fold CV within $m \times 2$ BCV should be at least a weak classifier to improve the model performance by aggregation with m models.

Lemma 4. *When $m \to \infty$, the parameters $p_{1,m}$ and $p_{0,m}$ converge to the limits denoted by $p_{1,\infty}$ and $p_{0,\infty}$, which have the following forms,*

$$\begin{aligned} p_{1,\infty} &= \int_{0.5}^{1} Beta\Big(\frac{p_1(1-\rho_2)}{\rho_2}, \frac{(1-p_1)(1-\rho_2)}{\rho_2}\Big)dx, \\ p_{0,\infty} &= \int_{0}^{0.5} Beta\Big(\frac{(1-p_0)(1-\rho_2)}{\rho_2}, \frac{p_0(1-\rho_2)}{\rho_2}\Big)dx. \end{aligned} \tag{18}$$

Proof. As $m \to \infty$, the parameter m_e converge to m_∞, i.e., $m_\infty = \lim_{m\to\infty} m_e = \frac{1}{\rho_2}$. Then, the parameters $p_{1,m}$ and $p_{0,m}$ converge to the limits denoted by $p_{1,\infty}$ and $p_{0,\infty}$ respectively, which can be expressed as follows:

$$\begin{aligned} p_{1,\infty} = \lim_{m\to\infty} p_{1,m} &= \lim_{m\to\infty} \int_{0.5}^{1} Beta\big((m_e-1)p_1, (m_e-1)(1-p_1)\big)dx \\ &= \int_{0.5}^{1} Beta\Big(\frac{p_1(1-\rho_2)}{\rho_2}, \frac{(1-p_1)(1-\rho_2)}{\rho_2}\Big)dx, \\ p_{0,\infty} = \lim_{m\to\infty} p_{0,m} &= \lim_{m\to\infty} \int_{0}^{0.5} Beta\big((m_e-1)(1-p_0), (m_e-1)p_0\big)dx \\ &= \int_{0}^{0.5} Beta\Big(\frac{(1-p_0)(1-\rho_2)}{\rho_2}, \frac{p_0(1-\rho_2)}{\rho_2}\Big)dx. \end{aligned}$$

Remark 3. In order to clearly show the variations of $p_{1,m}$ and $p_{0,m}$ with an increasing of m, and the limit values $p_{1,\infty}$ and $p_{0,\infty}$ when m tends to infinity, some simulation experiment results are given in Table 2.

From Table 2, we can see that $p_{1,m}$ and $p_{0,m}$ are increasing with regard to m and reach the maximum values that are the limit values $p_{1,\infty}$ and $p_{0,\infty}$ when m tends to infinity. However, the limit values do not tend to 1 even though ρ_2 is small. In particular, when ρ_2 is relatively large, the limit values $p_{1,\infty}$ and $p_{0,\infty}$ are not much bigger than p_1 and p_0. This means that the prediction by majority voting based on $m \times 2$ BCV will be effective only in the situation of relatively small ρ_2 and p_1, p_0 slightly larger than 0.5.

According to [5], the components TP_m, FP_m, FN_m and TN_m of voting confusion matrix $\mathcal{H}_m(D_n)$ follow the binomial distributions.

$$\begin{aligned}
TP_m|TP_m + FN_m &\sim Binom(n\pi_1, p_{1,m}),\\
FN_m|TP_m + FN_m &\sim Binom(n\pi_1, 1-p_{1,m}),\\
TN_m|TN_m + FP_m &\sim Binom(n\pi_0, p_{0,m}),\\
FP_m|TN_m + FP_m &\sim Binom(n\pi_0, 1-p_{0,m}).
\end{aligned}$$

The monotonic increase of $p_{1,m}$ and $p_{0,m}$ in Lemma 3 ensures the following inequalities for the expectations and variances of TP_m, FP_m, FN_m, TN_m. The limit values $p_{1,\infty}$ and $p_{0,\infty}$ in Lemma 4 establish the following bounds for these expectations and variances.

Lemma 5. *If $0 < \rho_2 < 1$, then*

$$\begin{aligned}
&n\pi_1 p_{1,\infty} \geq E(TP_{m+1}) \geq E(TP_m), \quad && n\pi_0(1-p_{0,\infty}) \leq E(FP_{m+1}) \leq E(FP_m),\\
&n\pi_0 p_{0,\infty} \geq E(TN_{m+1}) \geq E(TN_m), && n\pi_1(1-p_{1,\infty}) \leq E(FN_{m+1}) \leq E(FN_m),
\end{aligned}$$

and

$$\begin{aligned}
n\pi_1 p_{1,\infty}(1-p_{1,\infty}) \leq Var(TP_{m+1}) \leq Var(TP_m),\\
n\pi_0 p_{0,\infty}(1-p_{0,\infty}) \leq Var(FP_{m+1}) \leq Var(FP_m),\\
n\pi_1 p_{1,\infty}(1-p_{1,\infty}) \leq Var(FN_{m+1}) \leq Var(FN_m),\\
n\pi_0 p_{0,\infty}(1-p_{0,\infty}) \leq Var(TN_{m+1}) \leq Var(TN_m),
\end{aligned} \tag{19}$$

where expectations $E(\cdot)$ and variances $Var(\cdot)$ are taken for any dataset D_n with size n.

Proof. Recall that the components TP_m, FP_m, FN_m, and TN_m follow the binomial distribution given the counts $TP_m + FN_m \equiv n\pi_1$ and $TN_m + FP_m \equiv n\pi_0$, that is,

$$\begin{aligned}
TP_m|TP_m + FN_m &\sim Binom(n\pi_1, p_{1,m}),\\
FN_m|TP_m + FN_m &\sim Binom(n\pi_1, 1-p_{1,m}),\\
TN_m|TN_m + FP_m &\sim Binom(n\pi_0, p_{0,m}),\\
FP_m|TN_m + FP_m &\sim Binom(n\pi_0, 1-p_{0,m}).
\end{aligned}$$

Then, since the counts $m\pi_1$ and $n\pi_0$ are constants, we can derive the expectations and variances of these components as follows:

$$\begin{aligned}
E(TP_m) &= n\pi_1 p_{1,m}, && E(FN_m) = n\pi_1(1-p_{1,m}),\\
E(TN_m) &= n\pi_0 p_{0,m}, && E(FP_m) = n\pi_0(1-p_{0,m}),\\
Var(TP_m) &= n\pi_1 p_{1,m}(1-p_{1,m}), && Var(FN_m) = n\pi_1(1-p_{1,m})p_{1,m},\\
Var(TN_m) &= n\pi_0 p_{0,m}(1-p_{0,m}), && Var(FP_m) = n\pi_0(1-p_{0,m})p_{0,m}.
\end{aligned}$$

From Lemma 3 and 4, we know that $p_{1,m}$ and $p_{0,m}$ are increasing functions of m when $p_1 > 0.5, p_0 > 0.5$, and converge to $p_{1,\infty}$ and $p_{0,\infty}$ respectively

when $m \to \infty$. On this basis, we can obtain the following inequalities for the expectations:

$$n\pi_1 p_{1,\infty} \geq E(TP_{m+1}) \geq E(TP_m), \quad n\pi_0(1-p_{0,\infty}) \leq E(FP_{m+1}) \leq E(FP_m),$$
$$n\pi_0 p_{0,\infty} \geq E(TN_{m+1}) \geq E(TN_m), \quad n\pi_1(1-p_{1,\infty}) \leq E(FN_{m+1}) \leq E(FN_m).$$

Then, we can derive the following inequalities for the variances:

$$\begin{aligned} Var(TP_{m+1}) - Var(TP_m) &= n\pi_1 p_{1,m+1}(1-p_{1,m+1}) - n\pi_1 p_{1,m}(1-p_{1,m}) \\ &= n\pi_1(p_{1,m+1}-p_{1,m})(1-(p_{1,m}+p_{1,m})) < 0, \\ Var(FP_{m+1}) - Var(FP_m) < 0, &\qquad Var(TN_{m+1}) - Var(TN_m) < 0, \\ Var(FN_{m+1}) - Var(FN_m) < 0. & \end{aligned}$$

Finally, we can obtain the inequalities of variances in Eq. (19).

Furthermore, using the bounds of the expectations in Lemma 5 as the parameters of the beta distributions, we define the random variables $\tilde{p}$, $\tilde{r}$, and $\tilde{f}_1$, that is,

$$\begin{aligned} \tilde{p} \sim Beta(n\pi_1 p_{1,\infty}, n\pi_0(1-p_{0,\infty})), &\qquad \tilde{r} \sim Beta(n\pi_1 p_{1,\infty},\ n\pi_1(1-p_{1,\infty})), \\ \tilde{f}_1 \sim \frac{2^{a_\infty}(1-t)^{a_\infty-1}(2-t)^{-a_\infty-b_\infty}t^{b_\infty-1}}{B(a_\infty, b_\infty)}, & \end{aligned} \tag{20}$$

where $a_\infty = n\pi_0(1-p_{0,\infty}) + n\pi_1(1-p_{1,\infty})$ and $b_\infty = n\pi_1 p_{1,\infty}$.

Then, the modes and variances of the distributions of $\tilde{p}$, $\tilde{r}$, and $\tilde{f}_1$ can be defined as:

$$mode(\tilde{p}) = P_\infty = \frac{\pi_1 p_{1,\infty}}{\pi_1 p_{1,\infty} + \pi_0(1-p_{0,\infty})},\ Var(\tilde{p}) = \frac{\pi_1 p_{1,\infty}\cdot\pi_0(1-p_{0,\infty})}{n(\pi_1 p_{1,\infty} + \pi_0(1-p_{0,\infty}))^3},$$
$$mode(\tilde{r}) = R_\infty = p_{1,\infty}, \qquad Var(\tilde{r}) = \frac{p_{1,\infty}(1-p_{1,\infty})}{n\pi_1},$$
$$mode(\tilde{f}_1) = F_{1,\infty}, \qquad Var(\tilde{f}_1) = Var(f_1|\lambda=1, a=a_\infty, b=b_\infty).$$

For the random variable $\tilde{p}$, the mode P_∞ and variance $Var(\tilde{p})$ can be regarded as the limit values of the expectations both the mode P_m and variance $Var(p|\mathcal{H}_m(D_n))$ when m tends to infinity. For the random variables $\tilde{r}$ and $\tilde{f}_1$, the similar results can be obtained.

$$P_\infty = \lim_{m\to\infty} E_{D_n}(P_m), \quad Var(\tilde{p}) = \lim_{m\to\infty} E_{D_n}(Var(p|\mathcal{H}_m(D_n))). \tag{21}$$

D Proof of Theorem 1

Theorem 1. *If $0 < \rho_2 < 1$, we have*

$$\begin{aligned} &P_\infty > E(P_{m+1}) > E(P_m) > E(P_a), \\ &R_\infty > E(R_{m+1}) > E(R_m) > E(R_a), \\ &F_{1,\infty} > E(F_{1,m+1}) > E(F_{1,m}) > E(F_{1,a}), \end{aligned} \tag{22}$$

where expectations $E(\cdot)$ are taken with respect to dataset D_n of size n. Furthermore, as long as the correlation coefficient ρ_2 is not close to zero, the upper bounds P_∞,R_∞, and $F_{1,\infty}$ will not be close to one.

Proof. **[1] The monotonic increasing property of expectations of the proposed voting aggregation estimators of p, r, and f_1 measure.**

Using Taylor expansion at the points $E(TP_m)$ and $E(FP_m)$ for P_m, we have

$$\begin{aligned} P_m \approx & \frac{E(TP_m)}{E(TP_m)+E(FP_m)} + \frac{\partial P_m}{\partial TP_m}(TP_m - E(TP_m)) \\ & + \frac{\partial P_m}{\partial FP_m}(FP_m - E(FP_m)), \end{aligned}$$

and

$$E(P_m) \approx \frac{E(TP_m)}{E(TP_m)+E(FP_m)} = \frac{1}{1+\frac{E(FP_m)}{E(TP_m)}}.$$

Then, from the monotonic increasing property of $E(TP_m)$ and the monotonic decreasing property of $E(FP_m)$ with an increasing of m shown in Lemma 5, we can obtain that $E(P_{m+1}) > E(P_m)$.

For $E(R_m)$, according to the property of confusion matrix, we have

$$\begin{aligned} E(R_{m+1}) - E(R_m) &= E\Big(\frac{TP_{m+1}}{TP_{m+1}+FN_{m+1}}\Big) - E\Big(\frac{TP_m}{TP_m+FN_m}\Big) \\ &= \frac{E(TP_{m+1}-TP_m)}{n\pi_1} > 0. \end{aligned}$$

That is, $E(R_{m+1}) > E(R_m)$.

For $E(F_{1,m})$, recalling the mode (i.e., Eq. (3)) of the posterior distribution of f_1 measure in Section 2, we have

$$\begin{aligned} F_{1,m} =& mode(f_1|\mathcal{H}_m(D_n)) \\ \equiv& 1.25 - 0.25a - 0.5b + 0.25\sqrt{4b^2+4ab-4b+a^2-10a+9} \\ =& 1 - 0.25(a+2b-1) + 0.25\big[(a+2b-1)^2 - 8(a-1)\big]^{0.5}, \end{aligned}$$

where $a \approx FP_m + FN_m$ and $b \approx TP_m$. Using Taylor's first-order expansion at the points $E(TP_m)$, $E(FP_m)$, and $E(FN_m)$, we have

$$\begin{aligned} E(F_{1,m}) \approx & 1 - 0.25\big(E(a) + 2E(b) - 1\big) \\ & + 0.25\Big[\big(E(a)+2E(b)-1\big)^2 - 8\big(E(a)-1\big)\Big]^{0.5}. \end{aligned}$$

Denoted as $g(x,y) \equiv E(F_{1,m})$, $x \equiv E(a)$, and $y \equiv E(b)$, we have

$$g(x,y) \approx 1 - 0.25(x+2y-1) + 0.25[(x+2y-1)^2 - 8(x-1)]^{0.5}.$$

Taking partial derivatives for x and y in $g(x,y)$, respectively, we have

$$\begin{aligned}\frac{\partial g(x,y)}{\partial x} &= -0.25 + 0.25\frac{2(x+2y-1)-8}{2\sqrt{(x+2y-1)^2 - 8(x-1)}} \\ &= -0.25 + \frac{0.25(x+2y-1)-1}{\sqrt{(x+2y-1)^2 - 8(x-1)}} \\ &= -\frac{g(x,y)}{\sqrt{(x+2y-1)^2 - 8(x-1)}} \\ &= -\frac{g(x,y)}{x+2y+4g(x,y)-5},\end{aligned}$$

$$\begin{aligned}\frac{\partial g(x,y)}{\partial y} &= -0.5 + 0.25\frac{4(x+2y-1)}{2\sqrt{(x+2y-1)^2 - 8(x-1)}} \\ &= -0.5 + \frac{0.25(x+2y-1)}{0.5\sqrt{(x+2y-1)^2 - 8(x-1)}} \\ &= \frac{1-g(x,y)}{0.5\sqrt{(x+2y-1)^2 - 8(x-1)}} \\ &= \frac{1-g(x,y)}{0.5x+y+2g(x,y)-2.5}.\end{aligned}$$

According to $0 < g(x,y) < 1$ and $x, y > 0$, it is easy to obtain the following formulas.

$$\begin{aligned}\frac{\partial g(x,y)}{\partial x} &= -\frac{g(x,y)}{x+2y+4g(x,y)-5} < 0, \\ \frac{\partial g(x,y)}{\partial y} &= \frac{1-g(x,y)}{0.5x+y+2g(x,y)-2.5} > 0.\end{aligned}$$

From these formulas, we know that $g(x,y)$ is a monotonically decreasing function of x, and is an increasing function of y. Thus, $E(F_{1,m})$ is a monotonically decreasing function of $E(FP_m + FN_m)$, and is a monotonically increasing function of $E(TP_m)$. Furthermore, combining the former one with the property that $E(FP_m + FN_m)$ is a decreasing function of m, we can conclude that $E(F_{1,m})$ is a monotonically increasing function of m. Similarly, combining the latter one with the property that $E(TP_m)$ is an increasing function of m, we also conclude that $E(F_{1,m})$ is a monotonically increasing function of m. Therefore, we can obtain that

$$E(F_{1,m+1}) > E(F_{1,m}).$$

[2] The upper bounds of expectations of the proposed voting aggregation estimators of p, r, and f_1 measure.

First, from the Lemma 5, the upper bound of $E(P_{m+1})$ is obviously P_∞. Then, for $E(R_m)$, from the Lemma 5, we can easily obtain that

$$E(R_m) = E\Big(\frac{TP_m}{TP_m + FN_m}\Big) = p_{1,m} < p_{1,m+1} < p_{1,\infty}.$$

Therefore, the upper bound of $E(R_{m+1})$ is R_∞. Similarly, the upper bound of $E(F_{1,m})$ is $F_{1,\infty}$.

[3] The voting aggregation estimators of p, r, and f_1 measure are larger than average estimators of p, r, and f_1 measure.

Recall that the average confusion matrix $\mathcal{H}_a(D_n)$ based on $m \times 2$ BCV is defined as follows.

$$\begin{aligned}\mathcal{H}_a(D_n) &= (TP_a, FP_a, FN_a, TN_a)\\ &= \frac{\omega}{2m}\sum_{i=1}^{m}\sum_{k=1}^{2}(TP_k^{(i)}, FP_k^{(i)}, FN_k^{(i)}, TN_k^{(i)}).\end{aligned}$$

Notice that $m \times 2$ BCV is a special version of $m \times 2$ CV with some regularization conditions, including the identical distributions of training and validation data in each two-fold CV and the approximately equal numbers of overlapping samples between the training sets (validation sets) for any two two-fold CV. These conditions guarantee that each confusion matrix $\mathcal{H}_k^{(i)} = (TP_k^{(i)}, FP_k^{(i)}, FN_k^{(i)}, TN_k^{(i)})$ in the collection $\{\mathcal{H}_1^{(i)}, \mathcal{H}_2^{(i)}\}_{i=1}^m$ from $m \times 2$ BCV have the same expectations for the components $TP_k^{(i)}$, $FP_k^{(i)}$, $FN_k^{(i)}$, and $TN_k^{(i)}$, which is independent to k and i.

Therefore, according to the property of confusion matrix, the expectations of components TP_a, FP_a, FN_a, and TN_a can be obtained.

$$\begin{aligned}E(TP_a) &= \omega\frac{n}{2}\pi_1 p_1 = \omega E(TP_k^{(i)}),\\ E(FP_a) &= \omega\frac{n}{2}\pi_0(1-p_0) = \omega E(FP_k^{(i)}),\\ E(TN_a) &= \omega\frac{n}{2}\pi_0 p_0 = \omega E(TN_k^{(i)}),\\ E(FN_a) &= \omega\frac{n}{2}\pi_1(1-p_1) = \omega E(FN_k^{(i)}).\end{aligned}$$

Similar to the P_m and R_m, the expectations of P_a and R_a are defined.

$$\begin{aligned}E(P_a) &\approx \frac{E(TP_a)}{E(TP_a) + E(FP_a)} = \frac{E(TP_k^{(i)})}{E(TP_k^{(i)}) + E(FP_k^{(i)})},\\ E(R_a) &\approx \frac{E(TP_a)}{E(TP_a) + E(FN_a)} = \frac{E(TP_k^{(i)})}{E(TP_k^{(i)}) + E(FN_k^{(i)})}.\end{aligned} \tag{23}$$

Note that the expectations of P_a and R_a have not related to ω, and ω only affects the estimated variance. Moreover, each matrix $(\mathcal{H}_1^{(i)}, \mathcal{H}_2^{(i)})$ in the collection $\{(\mathcal{H}_1^{(i)}, \mathcal{H}_2^{(i)})\}_{i=1}^m$ can be regarded as the voting confusion matrix in situation

of $m = 1$, and the components TP_a, FP_a, FN_a, TN_a in the average confusion matrix $\mathcal{H}_a(D_n)$ neither increase nor decrease with the changes of m. Therefore, the average estimators P_a, R_a, and $F_{1,a}$ can be seen as the voting aggregation estimators with $m = 1$ and $\omega = 2$. Thus, when $m > 1$, we have

$$E(P_m) > E(P_a),\ E(R_m) > E(R_a),\ E(F_{1,m}) > E(F_{1,a}). \tag{24}$$

E Proof of Theorem 2

Theorem 2. *When null hypothesis H_0 holds, the probability $P(H_0)$ and the Bayes factor $BF(\nu)$ satisfy the following lower bounds:*

$$P(H_0) \geq sigmoid(SNR^2(\Delta\nu)), \qquad BF(\nu) = \frac{P(H_0)}{P(H_1)} \geq SNR^2(\Delta\nu). \tag{25}$$

When alternative hypothesis H_1 holds, the probability $P(H_1)$ and the Bayes factor $BF(\nu)$ satisfy the following bounds:

$$P(H_1) \geq sigmoid(SNR^2(\Delta\nu)), \qquad BF(\nu) = \frac{P(H_0)}{P(H_1)} \leq \frac{1}{SNR^2(\Delta\nu)}, \tag{26}$$

where $SNR(\Delta\nu)$ is the signal-to-noise ratio of $\Delta\nu$.

Proof. **[1] Supposing that H_0 holds, the proofs are as follows.**

As m tends to infinity, according to the one-sided Chebyshev inequality, we have

$$\begin{aligned}
P(H_0) &= P\big(\nu_n^{\mathcal{A}} > \nu_n^{\mathcal{B}} \big| \mathcal{H}_m^{\mathcal{A}}(D_n), \mathcal{H}_m^{\mathcal{B}}(D_n)\big) \\
&= P\big((\nu_n^{\mathcal{A}} - \nu_n^{\mathcal{B}}) - E(\nu_n^{\mathcal{A}} - \nu_n^{\mathcal{B}}) > E(\nu_n^{\mathcal{B}} - \nu_n^{\mathcal{A}})\big) \\
&\geq \frac{E^2(\nu_n^{\mathcal{A}} - \nu_n^{\mathcal{B}})}{Var(\nu_n^{\mathcal{A}} - \nu_n^{\mathcal{B}}) + E^2(\nu_n^{\mathcal{A}} - \nu_n^{\mathcal{B}})} \\
&\geq \frac{SNR^2(\Delta\nu)}{1 + SNR^2(\Delta\nu)} = sigmoid(SNR^2(\Delta\nu)). \\
&\Rightarrow \big(1 + SNR^2(\Delta\nu)\big)P(H_0) \geq SNR^2(\Delta\nu) \\
&\Rightarrow SNR^2(\Delta\nu)\big(1 - P(H_0)\big) \leq P(H_0) \\
&\Rightarrow BF(\nu) = \frac{P(H_0)}{P(H_1)} \geq SNR^2(\Delta\nu)
\end{aligned}$$

[2] Suppose H_1 holds, the proofs are as follows.

As m tends to infinity, according to the one-sided Chebyshev inequality, we have

$$\begin{aligned}
P(H_1) &= P\big(\nu_n^{\mathcal{A}} \leq \nu_n^{\mathcal{B}} \big| \mathcal{H}_m^{\mathcal{A}}(D_n), \mathcal{H}_m^{\mathcal{B}}(D_n)\big) \\
&= P\big((\nu_n^{\mathcal{A}} - \nu_n^{\mathcal{B}}) - E(\nu_n^{\mathcal{A}} - \nu_n^{\mathcal{B}}) \leq E(\nu_n^{\mathcal{B}} - \nu_n^{\mathcal{A}})\big) \\
&\geq \frac{E^2(\nu_n^{\mathcal{A}} - \nu_n^{\mathcal{B}})}{Var(\nu_n^{\mathcal{A}} - \nu_n^{\mathcal{B}}) + E^2(\nu_n^{\mathcal{A}} - \nu_n^{\mathcal{B}})} \\
&\geq \frac{SNR^2(\Delta\nu)}{1 + SNR^2(\Delta\nu)} = sigmoid(SNR^2(\Delta\nu)). \\
&\Rightarrow \big(1 + SNR^2(\Delta\nu)\big)P(H_1) \geq SNR^2(\Delta\nu) \\
&\Rightarrow \big(1 - P(H_1)\big)SNR^2(\Delta\nu) \leq P(H_1) \\
&\Rightarrow BF(\nu) = \frac{P(H_0)}{P(H_1)} \leq \frac{1}{SNR^2(\Delta\nu)}.
\end{aligned}$$

F The Detailed Information of Tasks and Datasets

SRL Task. The task of SRL is conducted with bidirectional long short-term memory (BiLSTM). The two hyper-parameter configurations GloVe 100 and Random 100 correspond to models $\mathcal{A}$ and $\mathcal{B}$, respectively. The corpus of Chinese FrameNet (CFN) is used as a population, and a text corpus D_n of size n is sampled from a population without replacement, where n is the number of semantic roles and is set to 78749.

NER Task. The task of NER is to identify the boundaries of all NER chunks without recognizing their types and turning them into a sequential tagging problem at the word level. Two tag sets "IOB2" and "IOBES" based on the conditional random fields (CRFs) are employed as two models $\mathcal{A}$ and $\mathcal{B}$, respectively. The CoNLL 2003 English NER corpus is used as a population and the size n of text corpus D_n is the number of NER chunks and is set to 11247.

NER-ORG Task. The task of NER-ORG is to identify only "ORG" entities in the CoNLL 2003 English NER training set. The text corpus and tag sets are similar to those of the NER task. The $m \times 2$ BCV is adopted for all tasks to split the whole text corpus and m is set to 15.